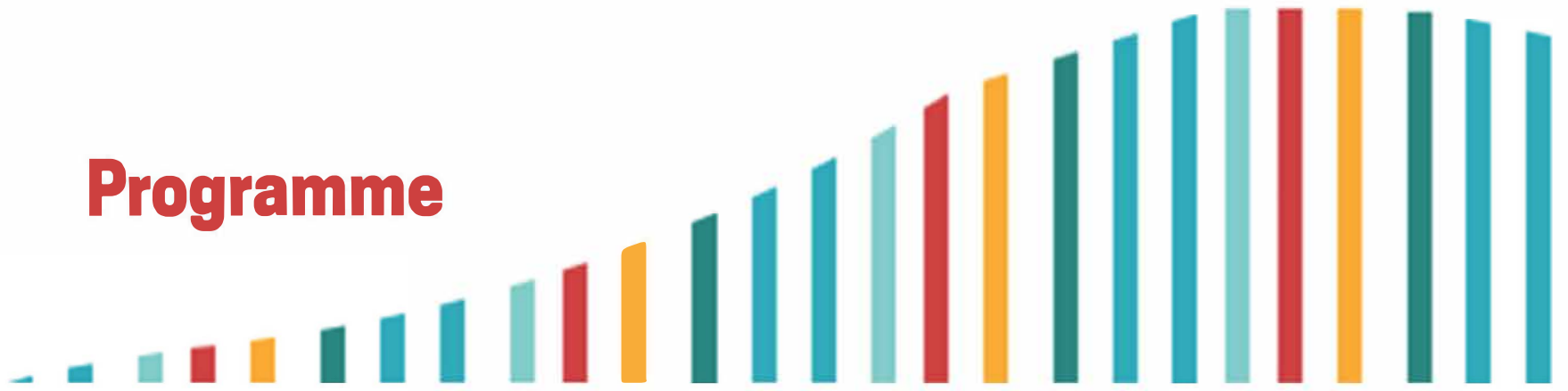


SASA 2017

Programme





BLUE STALLION TECHNOLOGIES cc

Technical, Scientific and Business Software Solutions
Visit us at www.bluestallion.co.za



Dear SASA2017 Conference Delegate

It is a great pleasure to welcome you all in the heart of the Free State, where the sky is bigger and the stars shine brighter. I trust the Free State's friendly people, casual manner, appropriate pace, lack of traffic and terrific food will make your stay unforgettable. Although we have a very full scientific programme, we managed to accommodate two excursions in the schedule, which is a first for SASA. The Free State also offers excellent golf courses, fine dining, a planetarium, and coffee shops to name only a few. Why not come earlier, or linger a little longer after the conference.



Packed with interesting talks, our scientific programme will make your selection of sessions to attend quite tough. We address this problem with another first: Our online web App (bloemsasa.co.za) provides easy navigation between different sessions and time slots.

Daily we hear buzzwords like machine learning, real time processing, data science, state capture, blockchain technology, bitcoin, BIG DATA and many others. I am sure we will learn much about these topics during the conference, if only because of the many networking opportunities.

The new goldmines
are data piles

Data nowadays are collected and stored, and sometimes made public at a faster pace than statisticians can analyse.

The new goldmines are data piles, the digging equipment our statistical software, using huge computing power. The mastermind statisticians who develop new models and techniques are those who can benefit, if they dare to dig. Our hope is that you leave our City of Roses with BIG methods, BIG networks and BIG ideas.

Frans F. Koning
Chair, LOC

General Information

Registration

The registration desk is situated in the foyer of the Ilanga Hall and will be open for registration all day. The preferred registration times are:

| | |
|-----------|---|
| Monday | 27 November 8:00 – 9:00 and 17:00 – 18:00 |
| Tuesday | 28 November 8:00 – 9:00 |
| Wednesday | 29 November 8:00 – 8:30 |
| Thursday | 30 November 8:00 – 8:30 |

All questions and queries should be directed to the staff at the registration desk (available from 8:00 – 17:00).

Parking

Free, secure parking is available at Ilanga estate.

Name Tags

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

Emergency numbers

| | |
|--------------------------|------------|
| ER24 | 0800051051 |
| Bayswater Police station | 0514062605 |

Poster session

All the posters will be on display in the right-side area of the Ilanga Hall. Posters must be up before 8:30 on Tuesday 28 November, and remain in the venue until 14:00 on Thursday 30 November.

WiFi Instructions

WiFi is available to ensure all delegates can access email and the conference app. While there are no restrictions on use, we request that delegates limit use of heavy traffic applications such as YouTube.

Network ID: SASA2017

Password: SASA2017

[Conference App - bloemsasa.co.za](http://bloemsasa.co.za)

The recommended way to view the programme is via the conference app. The app should work on any device and has the following features of note:

- a) Jumping to the current point in the programme. See the bottom right corner for this button.
- b) Navigating to adjacent time slots (Next Session at the bottom) or adjacent venues (Next Venue on the right).
- c) Searching for a time slot or speaker.
- d) Marking sessions you most want to see (Add to Favourites) so that you can find them quickly at any time (View Favourites).
- e) Announcements will be given via the app. The Announcements button will glow **green** if you have unread announcements.
- f) The ability to directly view the abstract of the presentation you are looking at or considering (no more searching for abstracts).
- g) A running photo gallery to which anybody may upload photos at any point. Please allow up to two hours for photos to be approved.
- h) The ability to rate sessions and provide anonymous feedback to presenters. Comments will be compiled and mailed to presenters, so use this feature to ask questions that you did not get a chance to during the presentation.
- i) Downloading of presentations and additional material, where enabled by the presenter.

Social Events

Welcoming Function: 27 November (Monday) at 18:00 in front of the Ilanga Hall.

Young Statisticians Function: 28 November (Tuesday) from 18:30 to 23:00 at Emoya Estate's Cheetah venue.

Cheetah Experience and Naval Hill Excursion: 29 November (Wednesday), busses leave at 13:15 from Ilanga Estate's parking area.

Gala Dinner: 29 November (Wednesday) 18:30 for 19:00 at Summerwood Estate.

Venues

Presentation venues: Ilanga Hall, Ruma I, Ruma II, Amanzi Hall and Chapel.

Poster venue: Righthand-side area of Ilanga Hall.

Tea/Coffee: Lawn in front of Ilanga Hall.

Lunch: Dinaledi Hall.

Exhibition Venue: Righthand-side area of Ilanga Hall.

Welcoming Function: Lawn in front of Ilanga Hall.

Gala dinner: Summerwood Estate.

Meetings

SASA Executive Committee meeting: 27 November (Monday) from 17:00 – 18:00 in Ilanga Hall.

SASA Annual General meeting: 28 November (Tuesday) from 14:40 – 15:40 in Ilanga Hall.

Biostatistics meeting: 28 November (Tuesday) from 10:50 – 11:20 in Amanzi Hall.

Data Science meeting: 28 November (Tuesday) from 15:40 – 16:10 in Amanzi Hall.

Multivariate Data Analysis Group (MDAG) meeting: 29 November (Wednesday) from 10:05 – 10:35 in Amanzi Hall.

Bayesian meeting: 30 November (Thursday) from 10:05 – 10:35 in Amanzi Hall.

Guidelines to Speakers and Chairpersons

Speakers:

Double check the date and time of your presentation.

Please submit your presentation electronically prior to travelling to the conference. If this is not possible, load your presentation between 8:00 and 8:30 on the computer at the registration desk in the foyer of Amanzi Hall.

Only standard presentation software will be available, including Adobe Acrobat and Microsoft Powerpoint. Please test your presentation on multiple platforms.

Report to the chairperson of the session before the start of the session.

Keep to the time allocated for your presentation.

Once the chair indicates the end of your session, you must stop your presentation immediately.

Chairpersons:

Double check the date and time of your session.

Keep to the scheduled times.

No changes are to be made to the programme.

Check the attendance of all the speakers, and ensure that all presentations have been loaded on the computer in the venue.

Welcome delegates and speakers at the beginning of your session.

Allow 15 minutes for each presentation and 5 minutes for questions.

Warn delegates visually 10 minutes into their talk that they have 5 minutes remaining.

Note that there is a 5 minute movement period between all slots after the question time. Please keep this time clear.

Thank all speakers and delegates at the end of the session.

Report any issues to the session assistant immediately.

Should you chair a double session you must still break for the 5 minute movement period.

Map to all venues



| Time | Monday | | | |
|-------|--|---|--|--|
| 00:00 | SASA2017 | | | |
| to | Ilanga Estate Bloemfontein | | | |
| 08:00 | South African Statistical Association | | | |
| | 59th Annual Conference | | | |
| 08:00 | Registration | | | |
| to | Amanzi Hall Foyer | | | |
| 09:00 | | | | |
| 09:00 | Workshop 1 | Workshop 2 | Workshop 3 | Workshop 4&5 |
| to | Chapel | Ruma I | Ruma II | Amanzi Hall |
| 10:30 | Emmanuel Lesaffre | Lixing Zhu | Hansjoerg Albrecher | Ulrich Paquet |
| | The analysis of interval-censored observations | Sufficient dimension reduction in regressions | Stochastic simulation and statistical modelling with applications in reinsurance | Variational methods in statistical inference |
| 10:30 | Morning Tea | | | |
| to | Lawn @ Ilanga Hall | | | |
| 11:00 | | | | |
| 11:00 | Workshop 1 | Workshop 2 | Workshop 3 | Workshop 4&5 |
| to | Chapel | Ruma I | Ruma II | Amanzi Hall |
| 12:30 | Emmanuel Lesaffre | Lixing Zhu | Hansjoerg Albrecher | Ulrich Paquet |
| | The analysis of interval-censored observations | Sufficient dimension reduction in regressions | Stochastic simulation and statistical modelling with applications in reinsurance | Variational methods in statistical inference |
| 12:30 | Lunch | | | |
| to | Dinaledi Hall | | | |
| 13:30 | | | | |

| Time | Monday | | | |
|-------|--|---|--|--|
| 13:30 | Workshop 1 | Workshop 2 | Workshop 3 | Workshop 4&5 |
| to | Chapel | Ruma I | Ruma II | Amanzi Hall |
| 15:00 | Emmanuel Lesaffre | Lixing Zhu | Hansjoerg Albrecher | Paul Kent |
| | The analysis of interval-censored observations | Sufficient dimension reduction in regressions | Stochastic simulation and statistical modelling with applications in reinsurance | SAS tools and SAS & Open source tools for the 3 things |
| 15:00 | Afternoon Tea | | | |
| to | Lawn @ Ilanga Hall | | | |
| 15:30 | | | | |
| 15:30 | Workshop 1 | Workshop 2 | Workshop 3 | Workshop 4&5 |
| to | Chapel | Ruma I | Ruma II | Amanzi Hall |
| 17:00 | Emmanuel Lesaffre | Lixing Zhu | Hansjoerg Albrecher | Paul Kent |
| | The analysis of interval-censored observations | Sufficient dimension reduction in regressions | Stochastic simulation and statistical modelling with applications in reinsurance | SAS tools and SAS & Open source tools for the 3 things |
| 17:00 | Registration | | | Executive Committee Meeting |
| to | Amanzi Hall Foyer | | | Ilanga Hall |
| 18:00 | | | | |
| 18:00 | Welcoming Function | | | |
| to | Lawn @ Ilanga Hall | | | |
| 20:00 | | | | |
| | Live music by Arco Musica | | | |



| Time | Tuesday |
|-------|---|
| 08:00 | Registration |
| to | Amanzi Hall Foyer |
| 08:30 | |
| 08:30 | Opening Ceremony |
| to | Ilanga Hall |
| 10:10 | Welcoming by SASA President Opening by LOC Chair Presidential address Statistician General's address Awards Platinum Sponsor address |
| 10:10 | Plenary 1 |
| to | Ilanga Hall |
| 10:50 | Paul Kent |
| | 3 things you need to do well to get full value from SAS analytics and SAS & Open source analytics |
| 10:50 | Tea / Biostatistics Meeting |
| to | Lawn @ Ilanga Hall |
| 11:20 | Meeting is in Amanzi Hall |

| Time | Tuesday | | | | |
|--------|--|--|--|---|--|
| Chair: | Ariane Neethling | Jaco Visagie | Edmore Ranganai | Mardi Jankowitz | Robert Schall |
| 11:20 | Business Statistics | Young Statisticians Competition | Theoretical Statistics | Stochastic Processes | Biostatistics |
| to | Ilanga Hall | Ruma I | Ruma II | Amanzi Hall | Chapel |
| 11:40 | Paul Kent | Gaonyalelwe Maribe | Njoku Ola Ama | Maxim Finkelstein | Divan Aristo Burger |
| | Data Lakes - is this really any different than the data warehouse we were building at the turn of the century. | Penalized bias reduction in extreme value estimation for censored Pareto-type data, and long-tailed insurance applications | Models for interaction in two-contingency tables | Optimal mission duration for partially repairable systems operating in a random environment | A Bayesian nonlinear mixed effects regression model for zero inflated negative binomial data, with application to the analysis of extended bactericidal activity of tuberculosis drugs |
| 11:45 | Business Statistics | Young Statisticians Competition | Theoretical Statistics | Stochastic Processes | Biostatistics |
| to | Ilanga Hall | Ruma I | Ruma II | Amanzi Hall | Chapel |
| 12:05 | Zukisa Mbava | Raeesa Ganey | Francois van Graan | Nil Kamal Hazra | Chris Muller |
| | A robust approach to assessing size and value of the JSE | Biplots based on Principal Surfaces | Smoothing Parameter Selection for Distribution Function Estimation using the Bootstrap | On a Hazard Rate Processes with Delays after Shocks | Dealing with missing data when modelling multi-state panel data |
| 12:10 | Plenary 2 | | | | |
| to | Ilanga Hall | | | | |
| 13:00 | Ulrich Paquet | | | | |
| | Statistics, everywhere | | | | |
| 13:00 | Lunch / Supervision Roundtable | | | | |
| to | Dinaledi Hall | | | | |
| 14:00 | Roundtable is next to Dinaledi Hall | | | | |
| | Porch on the Ilanga Hall side | | | | |
| 14:00 | Plenary 3 | | | | |
| to | Ilanga Hall | | | | |
| 14:40 | Hansjoerg Albrecher | | | | |
| | Modeling with very heavy tails and insurance of natural catastrophes | | | | |

| Time | Tuesday | | | | |
|--------|---|---|--|--|--|
| 14:40 | SASA AGM | | | | |
| to | Ilanga Hall | | | | |
| 15:40 | Data Science Meeting / Afternoon Tea | | | | |
| to | Lawn @ Ilanga Hall | | | | |
| 16:10 | Meeting is in Amanzi Hall | | | | |
| Chair: | Surette Bierman | WD (Willem) Schutte | Caston Sigauke | Njoku Ola Ama | Gary Sharp |
| 16:10 | Young Statisticians Competition | Business Statistics | General & Applied Statistics | Official Statistics | Young Statisticians Competition |
| to | Ilanga Hall | Ruma I | Ruma II | Amanzi Hall | Chapel |
| 16:30 | Carel van Niekerk | Michelle MacDevette | Warren Brettenny | Hellen Maribe | Danielle Roberts |
| | Feature detection using direct sampling and the Discrete Pulse Transform | Customer churn and fraud prediction in SPSS Modeler | The m-out-of-n Bootstrap for Inference in DEA Efficiency Assessments: An Application to South African Universities | A Quantitative Assessment of Municipal Revenue Sources in the Eastern Cape | The burden of childhood anaemia in four sub-Saharan African countries |
| 16:35 | Young Statisticians Competition | | General & Applied Statistics | Official Statistics | Young Statisticians Competition |
| to | Ilanga Hall | | Ruma II | Amanzi Hall | Chapel |
| 16:55 | Ané Neethling | | Peter Iiyambo | Arulsivanathan Naidoo | Albert Antwi |
| | Threshold autoregressive (TAR) time series models | | Coverage probabilities and average lengths of fiducial generalized confidence intervals for model parameters and quantiles of the Generalized Extreme Value distribution | Space time pattern mining of matrix pass rates | Exchange rate dependencies under the joint ARMAX-TGARCH System: Empirical evidence from the South African Forex Market |
| 17:00 | Young Statisticians Competition | General & Applied Statistics | General & Applied Statistics | Official Statistics | Young Statisticians Competition |
| to | Ilanga Hall | Ruma I | Ruma II | Amanzi Hall | Chapel |
| 17:20 | Johannes Ferreira | Allan Clark | Adenike Soogun | Thanyani Maremba | Liesl van Biljon |
| | Rank one noncentrality of complex Wishart type and its eigenvalue distributions | An alternate formulation of the occupancy model | Spatial Correlation of Diabetes across African Countries using Meta-Data | Complex survey sampling methodology applied to the South African Census 2011 small area layer data | A practical maturity assessment method for model risk management in banks |
| 18:30 | Young Statisticians Function | | | | |
| to | Emoya Estate Cheetah Hall | | | | |
| 23:00 | Karaoke | | | | |

| Time | Wednesday | | | | |
|--------|--|---|--|---|---|
| 08:00 | Registration | | | | |
| to | Amanzi Hall Foyer | | | | |
| 08:30 | | | | | |
| Chair: | Delia North | Leonard Santana | Roelof Coetzer | Michael von Maltitz | Renette Blignaut |
| 08:30 | Statistics in Education | Young Statisticians Competition | General & Applied Statistics | Multivariate Data Analysis | Biostatistics |
| to | Ilanga Hall | Ruma I | Ruma II | Amanzi Hall | Chapel |
| 08:50 | Pauline Masemola | Ansie Smit | Roelof Coetzer | Sean van der Merwe | Ruffin Mutambayi |
| | Overall Performance of Grade 12 Learners in the Eastern Cape and Mpumalanga Province | Extreme Distributions based on Process Characteristics | Data-driven discriminant analysis for solving complex problems in industry | A method for Bayesian regression modelling of composition data | A Comparative Investigation in the Analysis of Malaria Re-Infected Patients using Accelerated Failure Time Models and Cox Proportional Hazards Models |
| 08:55 | Statistics in Education | Young Statisticians Competition | General & Applied Statistics | Multivariate Data Analysis | Biostatistics |
| to | Ilanga Hall | Ruma I | Ruma II | Amanzi Hall | Chapel |
| 09:15 | Delia North | Zolisa Bleki | Jupiter Simbeye | Michael von Maltitz | Khehla Daniel Moloji |
| | | Laplace Variational Inference of the Single-season Site Occupancy Model using a Probit Link Function | Spatial effects on modelling of individual HIV status in Malawi: A generalized additive model approach | The Assessment of Multiple Imputation | Survival Analysis of HIV/AIDS Patients in the Limpopo Province, South Africa |
| 09:20 | Preparing High School Learners for a Career in Statistics through Appreciation of Basic Data Analytics | Young Statisticians Competition | General & Applied Statistics | Multivariate Data Analysis | Biostatistics |
| to | | Ruma I | Ruma II | Amanzi Hall | Chapel |
| 09:40 | | Priyanka Nagar | Jean-Claude Malela Majika | Wallina Oosthuizen | Renette Blignaut |
| | | Contributions to κ - μ models in wireless systems | Design of a side-sensitive double sampling \bar{X} control chart for monitoring the location process parameter | Goodness-of-fit test for multivariate distributions based on the Mahalanobis distance | Identifying survival risk factors of HIV/AIDS patients on antiretroviral treatment |
| 09:45 | Statistics in Education | Young Statisticians Competition | General & Applied Statistics | Multivariate Data Analysis | Biostatistics |
| to | Ilanga Hall | Ruma I | Ruma II | Amanzi Hall | Chapel |
| 10:05 | Murray de Villiers | Marothi Peter Letsoalo | Abrie van der Merwe | David Hofmeyr | Freedom Gumedze |
| | Building Modern, Competitive Statistics Programs through industry collaboration (Highlighting the SAS Global Academics Programs) | Assessing Variance Components of Multilevel Models for Social Science Data: Application to Teenage Pregnancy Data | Bayesian testing for process capability indices | Interactive Clustering using Projections | A longitudinal COM-Poisson regression model with random effects |

| Time | Wednesday | | | | |
|--------|--|--|---|---|--|
| 10:05 | Tea / MDAG Meeting | | | | |
| to | Lawn @ Ilanga Hall | | | | |
| 10:35 | Meeting is in Amanzi Hall | | | | |
| 10:35 | Plenary 4 | | | | |
| to | Ilanga Hall | | | | |
| 11:25 | Lixing Zhu | | | | |
| Chair: | Eeva Rapoo | Aude Ines Mbonda Tiekwe | Inger Fabris-Rotelli | Niel Le Roux | Freedom Gumedze |
| 11:30 | Statistics in Education | Young Statisticians Competition | General & Applied Statistics | Multivariate Data Analysis | Biostatistics |
| to | Ilanga Hall | Ruma I | Ruma II | Amanzi Hall | Chapel |
| 11:50 | Moeketsi Simon Mosia | Gerrit Grobler | Nontembeko Dudeni-Tlhone | Pieter Schoonees | Robert Keli |
| | Applications of Statistical Learning Theory Towards Prediction of Student Success | A pure jump interest rate model for the South African market. | A cost effective methodology for estimating and monitoring domestic solar usage patterns with an integrated data approach | An Assessment of Methods for Calculating Neural Reliability from EEG Recordings | Latent trait analysis of the effects of categorical covariates on binary outcomes using structural equation modeling |
| 11:55 | Statistics in Education | Young Statisticians Competition | General & Applied Statistics | Multivariate Data Analysis | Biostatistics |
| to | Ilanga Hall | Ruma I | Ruma II | Amanzi Hall | Chapel |
| 12:15 | Eeva Rapoo | Stefan Janse van Rensburg | Inger Fabris-Rotelli | Sugnet Lubbe | Esme Jordaan |
| | South African high school learners' understanding of research and statistics as a tool of research | A Skew-t-normal Generalised Autoregressive Score (STN-GAS) Model | A point process model for pulses of the DPT from an image | clusboot: An R package for visualising bootstrapping of cluster analysis | The use of Structural Equation Models to explore direct and indirect paths |
| 12:20 | Young Statisticians Competition | Young Statisticians Competition | General & Applied Statistics | Multivariate Data Analysis | Biostatistics |
| to | Ilanga Hall | Ruma I | Ruma II | Amanzi Hall | Chapel |
| 12:40 | Marion Delpont | Melkamu Dedefo Gishu | Ariane Neethling | Niel Le Roux | Ria Laubscher |
| | A spatial variant of the Gaussian mixture of regressions model | Spatiotemporal Mapping and Detection of Mortality Clusters Due to Cardiovascular Disease with Bayesian Hierarchical Framework Using Integrated Nested Laplace Approximation: Application in Kersa HDSS | Prediction Error Estimation of the Survey-Weighted Least Squares Model under Complex Sampling | Biplot visualisations of interactions in a bilinear model having a qualitative dependent variable | Recent approaches to analysing nutrition data |
| 12:45 | Young Statisticians Competition | Young Statisticians Competition | General & Applied Statistics | Multivariate Data Analysis | Biostatistics |
| to | Ilanga Hall | Ruma I | Ruma II | Amanzi Hall | Chapel |
| 13:05 | Murendeni Maurel Nemukula | Mark de Lancey | Bernard Moeketsi Hlalele | Morne Lamont | Nicolene Thiebaut |
| | A point process characterisation of extreme temperatures | Clustering big data | A regression analysis application to ecological drought vulnerability assessment: Eastern Free State, South Africa | Sparse linear discriminant analysis for high-dimensional data: Techniques and packages in R | Obtaining yield probabilities by using different CV%, for row/col and randomized block designs for different crops in cultivar selection |

| Time | Wednesday |
|-------|--|
| 13:05 | Lunch |
| to | Dinaledi Hall |
| 14:00 | Only for delegates not going on excursions. |
| 13:15 | Excursion |
| to | Cheetah Experience / Naval Hill |
| 15:00 | <p>Details: Delegates must reach correct bus by 13:15. Packed lunch will be provided on the bus. Busses will return in time for start of tea.</p> |
| | <div style="display: flex; justify-content: space-around; align-items: center;">   </div> |
| 15:00 | Afternoon Tea |
| to | Lawn @ Ilanga Hall |
| 15:30 | |

| Time | Wednesday | | | | |
|--------|--|---|---|---|--|
| Chair: | Chun-Sung Huang | Maseka Lesaoana | Esme R Jordaan | Sugnet Lubbe | Sean Van der Merwe |
| 15:30 | Business Statistics | Young Statisticians Competition | Biostatistics | Multivariate Data Analysis | General & Applied Statistics |
| to | Ilanga Hall | Ruma I | Ruma II | Amanzi Hall | Chapel |
| 15:50 | Chun-Kai Huang | Precious Mdlongwa | Alfred Stein | Luca Steyn | Chantelle Clohessy |
| | Filtered DPOT and PORT models for high quantile forecasting | The Kumaraswamy Log-Logistic Weibull Distribution: Model, Properties and Applications | Spatial statistical aspects in geo-health studies for neglected tropical diseases | Extreme Value-based Novelty Detection | Bayesian Tolerance Intervals for the Assessment of Energy Yield of a Photovoltaic System |
| 15:55 | Business Statistics | Young Statisticians Competition | Biostatistics | Multivariate Data Analysis | General & Applied Statistics |
| to | Ilanga Hall | Ruma I | Ruma II | Amanzi Hall | Chapel |
| 16:15 | Chun-Sung Huang | Bracken van Niekerk | Jacob Majakwara | Frances Coetzer | Elizabeth Girmay |
| | Forecasting Stock Market Volatility in the Presence of Structural Breaks: An Application to Value-at-Risk Estimation in South Africa | The Role of the Turning Angle when Modelling Animal Movement Data using Hidden Markov Models | Inference on the Destructive COM-Poisson Gamma Cure Rate Model | Aspects of Multi-Class Nearest Hypersphere Classification | Quantifying Eskom's Emissions Efficiency by Fitting The 3 Parameter Burr Distribution to the SO2 Emissions in Electricity Power Generation using Coal in South Africa. |
| 16:20 | Business Statistics | Young Statisticians Competition | Biostatistics | Multivariate Data Analysis | General & Applied Statistics |
| to | Ilanga Hall | Ruma I | Ruma II | Amanzi Hall | Chapel |
| 16:40 | Franck Adekambi | Adeboye Azeez | Innocent Karangwa | Yusentha Balakrishna | Norman Maswanganyi |
| | Second moment of the discounted aggregate renewal cash flow with dependence | Joint Modelling of Latent Class Mixed models for Longitudinal and Time-To-Event Outcomes | Providing insight on the WHO's HIV indicators for the population aged 15 – 24 years using the interval censoring time to event analysis: An empirical study using the 2012 South African National HIV Prevalence, HIV Incidence, Behaviour and Communication Survey | Statistical methods for the analysis of food consumption data | Peak electricity demand forecasting using partially linear additive quantile regression models |
| 16:45 | Business Statistics | Young Statisticians Competition | Biostatistics | Multivariate Data Analysis | General & Applied Statistics |
| to | Ilanga Hall | Ruma I | Ruma II | Amanzi Hall | Chapel |
| 17:05 | Harriet K Marima | Jesca Batidzirai | Tsirizani Kaombe | Oliver Bodhlyera | Caston Sigauke |
| | Risk management with long memory GARCH models: Empirical evidence from the USD/ZAR exchange rate | Analysis of Family Formation and Dissolution in Rural South Africa using Multi- State Transition Models | Detecting influential communities in the modelling of clustered child survival data in Malawi | Joint Modelling of the Evolution of Pulp Chemical Properties During Chemical Processing | Forecasting temporal hierarchical time series: An application to South African electricity data |
| 18:30 | Gala Dinner | | | | |
| to | Summerwood Estate | | | | |
| 23:00 | 18:30 for 19:00 | | | | |

| Time | Thursday | | | | |
|--------|--|--|---|---|---|
| 08:00 | Registration | | | | |
| to | Amanzi Hall Foyer | | | | |
| 08:30 | | | | | |
| Chair: | Jesca Batidzirai | Justin Harvey | James Allison | Victoria Goodall | Daniel Maposa |
| 08:30 | Young Statisticians Competition | Young Statisticians Competition | Theoretical Statistics | General & Applied Statistics | Time Series |
| to | Ilanga Hall | Ruma I | Ruma II | Amanzi Hall | Chapel |
| 08:50 | Jan Marais | Tshaudi Motsima | Theodor Loots | Stephan van der Westhuizen | Daniel Maposa |
| | Convolutional Neural Networks for Multi-Label Classification of Satellite Images | Determinants of Under-five mortality in South Africa: Results from a cross-sectional Demographic Health Survey | Approximations for the distribution of the product of independent beta random variables | The effect of unnecessary blocking on power in experimental design – a simulation study | Modelling nonstationary extremes in the lower Limpopo River basin of Mozambique |
| 08:55 | Young Statisticians Competition | Young Statisticians Competition | Theoretical Statistics | General & Applied Statistics | Statistical Software |
| to | Ilanga Hall | Ruma I | Ruma II | Amanzi Hall | Chapel |
| 09:15 | Johané Nienkemper-Swanepoel | Renate Thiede | Broderick Oluyede | Victoria Goodall | Clemens Dempers |
| | Competing approaches to the visualisation of incomplete categorical data sets | Road extraction in remote sensing images of South African informal settlements | A New Class of Log-logistic Modified Weibull Distributions with Applications | Crocodile Movements from the Olifants Gorge – insights from Hidden Markov Models | Applications of Wolfram Technology to Data Science |
| 09:20 | Young Statisticians Competition | Young Statisticians Competition | Theoretical Statistics | General & Applied Statistics | General & Applied Statistics |
| to | Ilanga Hall | Ruma I | Ruma II | Amanzi Hall | Chapel |
| 09:40 | Justine Naseje | Matthias Wagener | Not available | Lyness Matizirofa | Not available |
| | Application of Random Survival Forests in Understanding the Determinants of Under-Five Child Mortality in Uganda | On the Kumaraswamy-generalised normal distribution | Not available | Modelling predictors of stroke disease in South Africa: Bayesian binary quantile regression approach | Not available |
| 09:45 | Young Statisticians Competition | Young Statisticians Competition | Theoretical Statistics | General & Applied Statistics | General & Applied Statistics |
| to | Ilanga Hall | Ruma I | Ruma II | Amanzi Hall | Chapel |
| 10:05 | Sihle Poswayo | Nicola Gawler | Not available | Precious Mudavanhu | Sollie Millard |
| | Measuring Persistence in Time-Series using Paired Comparisons Judgements | Judicial decision-making: A statistical perspective | Not available | Improvement on the imputation model methodology for missing expenditure for the South African R&D survey data | An application of a mixture of logistic regressions |

| Time | Thursday | | | | |
|--------|--|--|---|---|---|
| 10:05 | Tea / Bayesian Meeting | | | | |
| to | Lawn @ Ilanga Hall | | | | |
| 10:35 | Meeting is in Amanzi Hall | | | | |
| Chair: | Paul Mostert | Nombuso Zondo | Hossein Masoumi Karakani | Reshoketswe Mokobane | Francois van Graan |
| 10:35 | Young Statisticians | Young Statisticians | General & Applied Statistics | General & Applied Statistics | Time Series |
| to | Ilanga Hall | Ruma I | Ruma II | Amanzi Hall | Chapel |
| 10:55 | Hayley Reynolds | Brett Rowland | Hossein Masoumi Karakani | Thapelo Sediadie | Tendai Makoni |
| | Spatial sampling scheme for a road network | Skew-normal distributions: Advances in theory and application | A Double Generally Weighted Moving Average Chart for Time Between Events | A Review On Survey Designs Adopted By Statistics Botswana | Modelling the volatility of international tourist arrivals in Zimbabwe using a GARCH process. |
| 11:00 | Young Statisticians | Young Statisticians | General & Applied Statistics | General & Applied Statistics | Time Series |
| to | Ilanga Hall | Ruma I | Ruma II | Amanzi Hall | Chapel |
| 11:20 | Carmen Stindt | Christine Papavarnavas | Prenil Sewmohan | David Rosevear | Sagaren Pillay |
| | Assessment of a composite index for analysing the goodness-of-fit in SEM | A LULU noise removal algorithm for images using principal component analysis | A comparison of the EM algorithm and the method of maximum likelihood under constraints | Gaussian Processes Applied to Class-Imbalanced Datasets | The impact of mis-specification of the deterministic trends on the cointegration vectors: an application to the bi-variate case on South African employment and earning |
| 11:25 | Plenary 5 | | | | |
| to | Ilanga Hall | | | | |
| 12:15 | Emmanuel Lesaffre | | | | |
| | Why interval censoring should not be ignored | | | | |
| 12:15 | Closing Ceremony | | | | |
| to | Ilanga Hall | | | | |
| 12:45 | | | | | |
| 12:45 | Packed Lunch | | | | |
| to | Dinaledi Hall | | | | |
| 13:30 | | | | | |
| 13:30 | End of conference | | | | |

Poster sessions

| Time | Presenter | Topic | Stream | Category |
|-----------------------------|------------------------|--|------------------------|----------|
| Tuesday 10:55 to 11:15 | Cindy Hayward | The application of natural language processing in the prediction of customer experience | Young Statisticians | Honours |
| Tuesday 10:55 to 11:15 | Motlamedi Thupae | Trees to networks: an evaluation of neural random forests | Young Statisticians | Honours |
| Tuesday 10:55 to 11:15 | Catherine Halsey | Minimum information for training a classifier | Young Statisticians | Honours |
| Tuesday 15:45 to 16:05 | Akhona Matshaya | Forecasting capacity loss in Eskom generation | Young Statisticians | Honours |
| Tuesday 15:45 to 16:05 | Siyabonga Mbonambi | A Bayesian Change Point Analysis of South African Financial Time Series | Young Statisticians | Honours |
| Tuesday 15:45 to 16:05 | Christopher Dunderdale | A Statistical Analysis of the Presence of Form in Cricket | Young Statisticians | Honours |
| Tuesday 15:45 to 16:05 | Michaela Laidlaw | Extension and alternative to the alpha-mu fading model | Theoretical Statistics | Honours |
| Wednesday 10:10 to 10:30 | Jacob Modiba | Sparse convex optimisation to solve a Sudoku problem | Young Statisticians | Masters |
| Wednesday 10:10 to 10:30 | Moshoko Emily Lebotsa | Short term load forecasting using quantile regression with an application to the unit commitment problem | Young Statisticians | Masters |
| Wednesday 10:10 to 10:30 | Simon Kashihalwa | Identifying Risk Factors to the Survival Time of Adult HIV Patients on Antiretroviral Therapy at Windhoek Public Hospitals, Namibia. A Retrospective Study with Proportional Hazard Model. | Biostatistics | Masters |
| Wednesday 10:10 to 10:30 | Jason Nakaluudhe | Modeling the impact of climate variability on diarrhoea-associated diseases in Zambia (2010-2013) | Biostatistics | Masters |
| Wednesday 10:10 to 10:30 | Onyekachi Esther Nwoko | Approaches for Handling Time-Varying Effects in Survival and Longitudinal Models | Biostatistics | Masters |

Poster sessions

| Time | Presenter | Topic | Stream | Category |
|-----------------------------|----------------------|---|--------------------------------------|----------|
| Wednesday 15:05 to 15:25 | Jani Deyzel | Assessment of Photovoltaic Energy Output using Bootstrap-based Tolerance Intervals | Applied Statistics and General | Masters |
| Wednesday 15:05 to 15:25 | Richard Southey | Bayesian Hierarchical Modelling with application in Spatial Epidemiology | Applied Statistics and General | Masters |
| Wednesday 15:05 to 15:25 | Raeesa Manjoo-Docrat | Generating Functions in Branching Processes and Birth and Death Processes | Time Series and Stochastic Processes | Masters |
| Wednesday 15:05 to 15:25 | Kirstie Eastwood | A Statistical Assessment of Solar Resource Data Across Multiple Sites in South Africa | Young Statisticians | Masters |
| Wednesday 15:05 to 15:25 | Peter Mphekgwana | Analysis of road traffic accidents in the Limpopo province using Generalized Linear modelling | Young Statisticians | Masters |
| Thursday 10:10 to 11:30 | Robert Schall | Three Statisticians in the Karoo (to say nothing of their wives) | Applied Statistics and General | Other |
| Thursday 10:10 to 11:30 | Robert Schall | Drie Statistici innie Kro (om van hun vrouwen nog maar te zwijgen) | Applied Statistics and General | Other |
| Thursday 10:10 to 11:30 | Siphamandla Sibiya | Bayesian Methods to Impute Missing Data from Climate Variables | Time Series and Stochastic Processes | Other |
| Thursday 10:10 to 11:30 | Pieter Barnard | Modelling industrial catalyst manufacturing using ensemble methods | Applied Statistics and General | Other |
| Thursday 10:10 to 11:30 | Ruffin Mutambayi | Polychotomous Logistic Regression Exploration of Determinants of Academic Achievement in the Lead-in Statistics Module: A Case Study of Undergraduate Students at the University of Fort Hare | Statistics in Education | Other |
| Thursday 10:10 to 11:30 | Peter Mphekgwana | Analysis of road traffic accidents in the Limpopo province using Generalized Linear modelling | Young Statisticians | Other |



ABSTRACTS



Second moment of the discounted aggregate renewal cash flow with dependence

Franck Adekambi - University of Johannesburg

We derive the second moment of the compound discounted renewal cash flow when taking into account dependence between the cash flow and its occurrence time. The dependence structure between the two random variables is defined by a Farlie-Gumbel-Morgenstern copula.

Keywords: Compound renewal model, Discounted aggregate cash flow, Moments, FGM copula, random interest rate

Coauthors:

Stream: Actuarial Science

Stochastic simulation and statistical modelling with applications in reinsurance

Hansjoerg Albrecher - Université de Lausanne

In this workshop, some variance reduction techniques in stochastic simulation will be discussed. Particular emphasis will be given on efficient rare-event simulation and simulation for dependent risks. We discuss modelling of losses and claim numbers as the main components of the insurance risk model, especially including challenges in models for long-tailed business lines. The techniques will be illustrated in reinsurance applications.

Keywords:

Coauthors:

Stream: Workshop

Modeling with very heavy tails and insurance of natural catastrophes

Hansjoerg Albrecher - Université de Lausanne

In this talk, some asymptotic properties of statistics based on very heavy tails are discussed, where the variance or even the mean of the underlying distribution does not exist. We then touch upon challenges and some recent developments of practical modeling approaches of flood and storm risk, for which such heavy tails play an important role. The approaches are illustrated in case studies based on actual loss data. Particular emphasis will be given on the choice of the underlying spatial and temporal dependence model.

Keywords:

Coauthors:

Stream: Plenary

Models for interaction in two-contingency tables

Njoku Ola Ama - University of South Africa

The author examines the similarity between the layout of the two factorial experiments with one observation per cell and the two way contingency table. Interaction in the two way contingency table is modelled in a manner similar to Tukey's (1949) and Johnson and Graybill (1972). One degree of freedom chi-square test statistics for interaction (independence) in the transformed and untransformed tables are developed. The criteria for transformation and choice of test statistics are established using the Asymptotic Relative Efficiency of the tests.

Keywords: Contingency table, Product effect model, Independence, Interaction

Coauthors:

Stream: Theoretical Statistics

Exchange rate dependencies under the joint ARMAX-TGARCH System: Empirical evidence from the South African Forex Market

Albert Antwi - University of Venda

The paper investigates the mean dependencies of the prices of CAD, BRL, USD, ILS and GBP on fourteen

traded currency pairs on the SA Forex Market under a system that jointly incorporates ARMAX and TGARCH processes. The model is linear in the mean and non-linear in the conditional variance; thus linear dependencies of conditional mean exchange rates on other exchange rates are studied under a non-linear process. Evidence from the study indicates that CAD prices significantly depend on the prices of AUD, MZN, NOK, ILS and NZD while BRL prices are significantly affected by the prices of CAD, AUD, NOK, NZD and USD. Except for BRL and INR, all other currency pairs in the GBP model have significant impact on the prices of GBP. USD prices significantly depend on all the currency pairs except for AUD, INR, KPW, MWK and EUR prices. ILS prices have significant impact on EUR, USD, BWP, NOK, NZD and GBP prices. All the significant dependencies are positive except for GBP on MWK, BRL on NZD and USD on NOK and NZD. The correlation among USD, NOK and NZD are positive but the impact of NOK and NZD on USD is negative. The outcome of the study suggests that correlation among currencies alone is not sufficient when selecting assets into a portfolio but also the impact of such correlations must be considered. The results also indicate that currency pairs can help in explaining the sources of volatility and improve volatility and pricing estimates when incorporated into such models.

Keywords: Exchange rates, dependencies, ARMAX-TGARCH

Coauthors: KA Kyei (University of Venda)

Stream: Financial and Business Statistics

Joint Modelling of Latent Class Mixed models for Longitudinal and Time-To-Event Outcomes

Adeboye Azeez - University of Fort Hare

In many clinical and reliability studies, the outcome of primary interest is the time to a particular event in order to indicate the individual's "true" state of health or survival status. To analyse such data, models have been developed to generate both longitudinal (repeated measurement) and survival (time to event) data separately, but these may be inappropriate when the longitudinal variable is correlated with patient health status. In this study, a latent class mixed model is used to model the correlation between covariates, unobserved latent class variables, status of health, and the observed variable indicators. The Cox model is extended to include latent class time-to-event outcomes. A Bayesian approach is employed to obtain maximum likelihood estimates implemented via Markov Chain Monte Carlo methods. The longitudinal and survival responses are assumed independent given a linking latent bivariate Gaussian process and the covariates. We use the approach to model jointly the longitudinal and survival data from a cardiovascular disease cohort study of the effect of severe mobility on time-to-event for adult with heart attack. Despite the complexity of the model, it is relatively straightforward to implement and understand, using the R package *hlme* *lcmm* *curvilinear* longitudinal outcomes and multivariate outcomes, as well as joint latent class mixed models (*Jointlcmm*) for a longitudinal outcome and a time-to-event.

Keywords: Cox model, EM algorithm, Gaussian, Multiple Regression, NonParametric, Random effect, Survival analysis

Coauthors: J Ndege (University of Fort Hare) and Y Qin (University of Fort Hare)

Stream: Biostatistics

Statistical methods for the analysis of food consumption data

Yusentha Balakrishna - South African Medical Research Council

A classical way to understand food choices is to combine consumption data (obtained from 24 hour/7 day recalls or quantified food frequency questionnaires) with a food composition database (FCDB). The resulting data set can be quite 'big' considering that FCDBs can contain over 7500 different foods and 140 food components. However, even though a large number of different foods are involved in individual dietary patterns, not all possible food combinations are observed in practice. One can then expect that the consumption data can be described by a few patterns of linear combinations of specific foods. Principal component analysis (PCA) and factor analysis are some of the methods currently used to extract such latent variables. We demonstrate the application of PCA to food consumption data and explore an alternative method for extracting dietary patterns.

Keywords: Nutrient intake, dietary patterns, principal component analysis

Coauthors: S Manda (South African Medical Research Council) and H Mwambi (University of KwaZulu-Natal)

Stream: Multivariate Data Analysis

Modelling industrial catalyst manufacturing using ensemble methods

Pieter Barnard - Sasol

Modelling the complex behaviour of industrial processes provides an elegant and tractable method of optimizing commercial plants without the need for costly experimentation. Catalyst manufacturing is the initial step in a chain of processes to produce wax. Therefore, it is crucial to optimize the quality of the catalyst to ensure sufficient supply and quality of the catalyst to subsequent plants. Due to the nature of the process and the changing conditions, data contains noise which adds to the complexity of modelling the process. This work describes ensemble methods to model the complexity of the catalyst manufacturing plant. Different ensemble approaches are tested on the data and their performance is compared to some popular machine learning algorithms. The final model is implemented in an online monitoring tool to provide the plant with the optimized process conditions and the ability to predict the catalyst quality.

Keywords: Ensemble models, Machine learning

Coauthors: RLJ Coetzer (Sasol)

Stream: Applied Statistics and General

Analysis of Family Formation and Dissolution in Rural South Africa using Multi- State Transition Models

Jesca Batidzirai - University of KwaZulu- Natal

The health, social and economic profiles of families and their children are strongly affected by the timing of marriage formation and dissolution. Using data from the Africa Health Research Institute, we estimate transition probabilities and transition intensities between marital states (never married, married, divorced, widowed and separated) and state durations through multi- state transition models and a competing risks discrete Markov process. The studied data contains 177 011 subjects who were enrolled between January 2000 and December 2016, some of whom were lost to follow- up. We considered HIV status, blood pressure, education level, employment status, income earned, gender and, where applicable, age at first marriage and age at first sex as predictors of various state transitions and probabilities.

Keywords: Transition Intensity; Transition Probability; Competing Risks; State Duration; Family Formation.

Coauthors: SO Manda (South African Medical Research Council), HG Mwambi (University of KwaZulu-Natal) and F Tanser (University of KwaZulu-Natal)

Stream: Biostatistics

Laplace Variational Inference of the Single-season Site Occupancy Model using a Probit Link Function

Zolisa Bleki - University of Cape Town

In the Bayesian context, Markov Chain Monte Carlo (MCMC) methods are powerful tools for deriving posterior distributions. However, when faced with large datasets and complex models, MCMC methods can be inefficient. In this paper we develop a Variational Bayes (VB) method using a probit link function in order to model posterior parameters of the single-season site occupancy model. We propose an efficient Gibbs sampler for modelling site occupancy data using Parameter-extended Data Augmentation (PX-DA). We show that the PX-DA method is more efficient than its Data Augmentation (DA) counterpart. We also show that the VB method can be used as a viable alternative to accurately model site occupancy data in situations where MCMC methods are computationally expensive and inefficient. To illustrate the VB method's performance a number of simulation settings are considered and the results are compared to those obtained by Maximum Likelihood Estimation and MCMC

methods. We show that the VB method can provide reliable estimates for datasets where the number of visits per site is as low as two, and that these estimates improve with an increase in sample size and number of visits. As an example, the VB method is used to analyse site occupancy data (obtained from the second Southern African Bird Atlas Project) of five different bird species.

Keywords: Detection Probability, Maximum Likelihood Estimation, Occupancy Modelling, Parameter-extended Data Augmentation, Species Occurrence, Variational Bayes

Coauthors: A Clark (University of Cape Town)

Stream: Young Statisticians

Identifying survival risk factors of HIV/AIDS patients on antiretroviral treatment

Renette Blignaut - University of the Western Cape

Purpose: The main aim of this study was to determine the risk factors that impact the survival of HIV/AIDS patients on ART. The study also seeks to highlight heterogeneity among patients from different clinical care facilities.

Methods: A retrospective cohort study design was used to collect information on adults who started ART between 2008 and 2010 at two hospitals in Windhoek, Namibia. A sample of 473 adults was randomly selected from the records of HIV/AIDS patients. Ethical approval to utilize the data was obtained from the Ministry of Health and Social Services, Namibia. Kaplan-Meier estimates, logrank tests and shared frailty proportional hazards models were used.

Results: Of the 473 patients, 62% were female. Most patients were less than 35 years (51%) and single (83%). The study shows starting ART with weight above 50kg was significantly protective against death and loss-to-follow up. The patients' baseline hazard varies significantly when hospital specific clinical procedures are included in the model. Survival estimates revealed that patients who started treatment in the early stages of the disease had a greater probability of survival compared to the more advanced stages. Baseline CD4 count, weight above 50kg and having a treatment supporter significantly influenced survival.

Keywords: Survival analysis, Kaplan-Meier, mortality, proportional hazards models, logrank tests

Coauthors: Innocent Maposa (Namibia University of Science and Technology)

Stream: Biostatistics

Joint Modelling of the Evolution of Pulp Chemical Properties During Chemical Processing

Oliver Bodhlyera - University of KwaZulu-Natal

This study looks at the statistical analysis of multivariate longitudinal data that involve several chemical properties of dissolving wood pulp. The study explains the interdependence of seven dissolving pulp properties, namely, viscosity, lignin, γ -cellulose, α -cellulose, copper number and glucose, as they evolve over the prescribed processing stages. The evolutionary correlations of these properties shed light on the dynamics of chemical pulp properties as they change over the processing stages. The results indicate that the evolution of viscosity is positively correlated to that of lignin ($r=0.6513$) and γ -cellulose ($r=0.7434$) and negatively correlated to that of α -cellulose ($r=-0.6401$) and, to a lesser extent, negatively correlated to the evolution of copper number ($r=-0.3788$) and glucose ($r=-0.4847$). Other relationships are also outlined. It is important to understand these correlations between chemical properties in order to identify properties that can be targeted jointly during processing, by varying the processing parameters like temperature and chemical concentrations.

Keywords: Chemical pulp properties, joint modelling, pairwise fitting

Coauthors: T Zewotir (University of KwaZulu-Natal) and S Ramroop (University of KwaZulu-Natal)

Stream: Multivariate Data Analysis

The m-out-of-n Bootstrap for Inference in DEA Efficiency Assessments: An Application to South African Universities

Warren Brettenny - Nelson Mandela University

The m-out-of-n bootstrap can be used to provide confidence interval estimation and bias correction in data envelopment analysis (DEA) efficiency models. The main benefit of this method is that it is free of the restrictive assumptions of the more popular homogeneous smoothed bootstrapping approach. Furthermore, the m-out-of-n bootstrap has the potential to be used for the detection of outliers in DEA assessments. The use of this method is investigated in an assessment of the efficiency of South African universities in both teaching and research activities.

Keywords: Efficiency, universities, bootstrap, DEA

Coauthors: A Gqwaka (Nelson Mandela University)

Stream: Applied Statistics and General

A Bayesian nonlinear mixed effects regression model for zero inflated negative binomial data, with

application to the analysis of extended bactericidal activity of tuberculosis drugs

Divan Aristo Burger - University of Pretoria

Extended bactericidal activity of tuberculosis drugs is conventionally assessed using "normal theory" regression modeling of logarithmically transformed colony forming unit counts (CFUs) over 8 weeks of treatment. In this study, a recently introduced Bayesian nonlinear mixed effects regression model was adapted to fit the data on the original count scale. Four probability distributions for CFU count were evaluated: Poisson, zero inflated Poisson, negative binomial, and zero inflated negative binomial. The deviance information criterion and compound Laplace-Metropolis marginal likelihoods were used to discriminate between the four candidate models, and a simulation study was carried out to compare their performance.

Keywords: Mixed effects model; zero inflated distribution; negative binomial distribution; extended bactericidal activity

Coauthors: R Schall (University of the Free State), D-G Chen (University of Pretoria and University of North Carolina - Chapel Hill) and R Jacobs (RIVM)

Stream: Biostatistics

An alternate formulation of the occupancy model

Allan Clark - University of Cape Town

Occupancy models (McKenzie et al, 2002) are widely used in ecology in order to develop species distribution maps for nature conservation purposes. We develop an alternate formulation of the (non-spatial and spatial) Bayesian model and use a simple Gibbs sampler to derive the posterior distribution of the parameters of the model. The methods are applied to data from the 2nd Southern African Bird Atlas Project database.

Keywords: Occupancy model; Bayesian analysis; Gibbs sampler

Coauthors:

Stream: Applied Statistics and General

Bayesian Tolerance Intervals for the Assessment of Energy Yield of a Photovoltaic System

Chantelle Clohessy - Nelson Mandela University

The growing interest in renewable energy has been driven by the negative impacts of fossil fuels on the environment, the volatility of fossil fuel prices and the need for national energy security. The renewable energy market is one of the fastest growing markets with energy being harnessed from a variety of abundant natural resources. These resources include solar, wind, geothermal, modern biomass and hydro. This study focusses its statistical assessment on solar energy and, in particular, the photovoltaic system. For investors to acquire finance to develop a photovoltaic system, an assessment of the forecasted energy yield and power output of the system is required. Typically, the energy yield and power output are assessed using basic descriptive statistics and the probability of exceedance. In this study, Bayesian tolerance intervals are proposed for the assessment of the power output and energy yield of a photovoltaic system. Tolerance intervals are used because of the similarities between process control in a manufacturing environment and energy generation from a photovoltaic system. The use of tolerance intervals is illustrated for a case study of a 1MW photovoltaic system proposed for the Nelson Mandela University Summerstrand South Campus in Port Elizabeth, South Africa.

Keywords: Bayesian Tolerance Intervals, Photovoltaic System, Energy Yield, Power Output

Coauthors: G Sharp (Nelson Mandela University), E van Dyk (Nelson Mandela University) and J Hugo (Nelson Mandela University)

Stream: Applied Statistics and General

Data-driven discriminant analysis for solving complex problems in industry

Roelof Coetzer - Sasol

In industry, production processes are not necessarily operated within design or at steady state. Furthermore, fundamental understanding and relationships are not always readily available for solving problems on a production plant. Therefore, data provide the only reliable information in complex processes, and a statistical modelling approach is required to investigate the relationships between many variables and to recommend

the optimal operating conditions for the performance indicator of interest. In this paper, the application of data-driven classification methods, such as regularized random forests, will be discussed for investigating the causes of corrosion in reactor lines and the identification of the most critical variables that affect corrosion. It will be shown how data-driven statistical analysis and modelling are applied in an innovative manner to solve a high-dimensional problem with direct business benefits.

Keywords: Data-driven analysis, Discriminant analysis, Random forests, Reactors

Coauthors: JA van den Berg (Sasol)

Stream: Applied Statistics and General

Aspects of Multi-Class Nearest Hypersphere Classification

Frances Coetzer - Stellenbosch University

Using hyperspheres in the analysis of multivariate data is not a common practice in statistics. However, hyperspheres have interesting properties which are useful for data analysis in the following areas: domain description (finding a support region), detecting outliers (novelty detection) and the classification of objects into known classes. In this talk hyperspheres are extended to multi-class classification, called nearest hypersphere classification (NHC). We need a dissimilarity function to classify objects into the nearest hypersphere. Different aspects of multi-class NHC are investigated. Using NHC requires choosing a kernel function, and we will use the Gaussian kernel. NHC also depends on selecting an appropriate kernel hyper parameter γ and a tuning parameter C . The behaviour of the error rate and the fraction of support vectors for different values of γ and C will be investigated. Two methods to obtain the optimal γ value for NHC were investigated. The first method uses a differential evolution procedure. The R function `DEoptim()` is used for this purpose. The second method uses the R function `sigest()`. The first method is dependent on the classification technique and the second method is executed independently of the classification technique. To study the classification performance of NHC we compared it to three other classification techniques: support vector machines, random forests and penalised LDA.

Keywords: Multi-class classification, nearest hypersphere classification, support vector machines, random forests, penalised LDA

Coauthors: F Coetzer (Stellenbosch University) and MMC Lamont (Stellenbosch University)

Stream: Multivariate Data Analysis

Constructing a Similarity/Dissimilarity Matrix from Spatial Series and Higher Order Lags for Use in the Generalised Regression Partial Mantel Test

Timotheus Brian Darikwa - University of Limpopo

The bivariate Mantel test was originally designed to detect spatio-temporal clustering in point pattern data. Its application has been extended to include, in general, any analysis relating to two similarity matrices. An extension of the bivariate Mantel test to multivariate spatial analysis has been carried out using partial correlation analysis between three or more similarity matrices and has been successfully applied. This paper proposes a new similarity matrix based on spatial series derived from binary weight matrices using univariate local indicators for use in extending Lee's and Moran's I bivariate spatial autocorrelations to multivariate spatial autocorrelations, for areal spatial data. Simulations were conducted to assess its usefulness. Results indicate that the new approach can be usefully implemented in multivariate spatial autocorrelation analysis of dependent health outcomes.

Keywords: Spatial series, Higher-order spatial lags, Multivariate spatial autocorrelation, Simple Mantel Test, Partial Mantel test, Local indicators of spatial autocorrelation, Heart disease

Coauthors: S Manda (South African Medical Research Council) and M Leasoana (University of Limpopo)

Stream: Applied Statistics and General

Clustering big data

Mark de Lancey - University of Pretoria

Clustering partitions data into meaningful groups or clusters and has important applications in artificial intelligence and pattern recognition. The CLARA algorithm is one of many clustering methods and is used for

clustering large sets of data. CLARA repeatedly samples a dataset and then applies the algorithm PAM, a k-medoids solver, to these samples and clusters the remainder of the data according to the medoids given by these sampled results. This report will examine the CLARA algorithm and how it can be used to cluster images, which are examples of big data, and compare it with other similar algorithms for clustering big data.

Keywords: cClustering, big data, image processing

Coauthors: I Fabris-Rotelli (University of Pretoria)

Stream: Young Statisticians

Clustering big data

Mark de Lancey - University of Pretoria

Clustering partitions data into meaningful groups or clusters and has important applications in artificial intelligence and pattern recognition. The CLARA algorithm is one of many clustering methods and is used for clustering large sets of data. CLARA repeatedly samples a dataset and then applies the algorithm PAM, a k-medoids solver, to these samples and clusters the remainder of the data according to the medoids given by these sampled results. This report will examine the CLARA algorithm and how it can be used to cluster images, which are examples of big data, and compare it with other similar algorithms for clustering big data.

Keywords: cClustering, big data, image processing

Coauthors: I Fabris-Rotelli (University of Pretoria)

Stream: Young Statisticians

Building Modern, Competitive Statistics Programs through industry collaboration (Highlighting the SAS Global Academics Programs)

Murray de Villiers - SAS

The ability to analyse data in real time and do predictive modelling enables faster, more informed decision-making, which allows businesses to better serve their customers, uncover new revenue streams, reduce risk and increase competitiveness. In-house data analytics skills are consequently in high demand in the business world, so that institutions of higher education are challenged to feed an ever increasing demand for more and more “job ready” statistics graduates. This talk will focus on collaborative projects between universities and industry, aimed at overcoming the problems outlined, thereby increasing the flow of “job ready” statisticians into the workplace.

Keywords: Data Analytics; curriculum design; job ready graduates

Coauthors:

Stream: Statistics in Education

Gaussian Processes Applied to Class-Imbalanced Datasets

Alta De Waal - University of Pretoria

Modelling class-imbalanced data is problematic. On such data, classifiers tend to misclassify minority class observations. Considering the potential practical use of a classifier that is particularly robust to class imbalance, the performance of a Gaussian process classifier is evaluated to determine the degree to which it addresses the problem.

Keywords: Gaussian processes, classification, class-imbalance

Coauthors: David Rosevear

Stream: Young Statisticians

A spatial variant of the Gaussian mixture of regressions model

Marion Delpert - Department of Statistics, University of Pretoria

Finite mixtures of Gaussian distributions and regressions assign observations to components based on a mixing probability, adding additional probabilistic information to the clustering result. Finite mixture models have been successfully applied in the image segmentation context but have been found lacking in that commonality of location of pixels is not taken into account when fitting the mixture models. The present paper discusses how this shortcoming can be addressed by imposing a Markov random field prior on the mixing probabilities. An application of the spatial variant finite mixture of regressions model in the agricultural context is contributed.

Keywords: Gaussian mixture of distributions, spatial, mixture of regressions, Markov random field, image segmentation

Coauthors: Kanfer, F.H.J. (Department of Statistics, University of Pretoria) and Millard, S.M. (Department of Statistics, University of Pretoria)

Stream: Applied Statistics and General

Applications of Wolfram Technology to Data Science

Clemens Dempers - Blue Stallion Technologies

Wolfram Mathematica and Wolfram Alpha are part of the Wolfram technology stack which is built on the uniquely flexible Wolfram Language. This toolset is well suited to applications in data science and analyzing and big data. Some of the key capabilities include: Integrated functionality extends to deep learning, neural networks and machine learning. A Vast range of statistical distributions, with specialized coverage for finance, medicine and science. Compatible with a wide range of data types, including images, text, sounds, GIS and numeric data. Automated time series model fitting and diagnostics. Systemwide support for random processes, including queueing, hidden Markov parametric processes and stochastic differential equation processes. Connectivity to the Hadoop network. The presentation will include live examples showcasing applications of Wolfram Mathematica, including how research results can be shared using the interactive computational document format (CDF).

Keywords:

Coauthors:

Stream: Special

Assessment of Photovoltaic Energy Output using Bootstrap-based Tolerance Intervals

Jani Deyzel - Nelson Mandela University

South Africa has many traditional and alternative energy resources. Renewable energy options, in particular solar energy, have been widely investigated and implemented. Currently, the “probability of exceedance” is the most popular method for assessing the energy output of a photovoltaic (PV) system. This study focuses on bootstrap-based tolerance intervals, traditionally used in a manufacturing environment, as an improved statistical assessment technique for the energy output of PV systems.

Keywords: Photovoltaic system; bootstrap; nonparametric; tolerance intervals.

Coauthors: C Clohessy (Nelson Mandela University), W Bretttenny (Nelson Mandela University) and E van Dyk (Nelson Mandela University)

Stream: Applied Statistics and General

Longitudinal Multilevel Models of Tuberculosis in South Africa: 2008 to 2014

Hilda Dhlakama - University of Johannesburg and University of KwaZulu Natal

Background: Although death due to Tuberculosis (TB) has been on the decline, TB was still the leading cause of death in 2013, accounting for 8.8% of all deaths in South Africa. Continued efforts to use research to create a nation free of TB are underway.

Aim: By determining the simultaneous multilevel and person-level predictors of TB, the aim of this research was to examine the categories of risk factors that are associated with TB infections and the extent to which these vary across TB patients belonging to the same province in South Africa over time for the period 2008 to 2014.

Method: A variety of Multilevel Models (MLM), Generalised Mixed Effects Models in particular, were applied to repeated measures of a sample taken from the National Income Dynamics Survey data that self-reported to have been TB diagnosed. The outcome variable was whether patients were on treatment or not. We used unbalanced wave data since MLMs can handle missing values.

Results: We found that over the years, males, the single, aged between 30 and 44, alcohol consumption, non-employment, English illiteracy, patients residing in houses with shared toilets and in informal dwelling types were highly associated with TB infections between 2008 and 2014.

Conclusion: Variation in TB infections is mainly at person level with behavioural risk factors contributing the most variation. Socio economic factors and poor housing conditions are also driving TB infections in South

Africa.

Keywords: Tuberculosis, Multilevel, Generalised Mixed Effects Models, South Africa, NIDS

Coauthors: S Lougue (University of KwaZulu-Natal)

Stream: Applied Statistics and General

A cost effective methodology for estimating and monitoring domestic solar usage patterns with an integrated data approach

Nontembeko Dudeni-Tlhone - Council for Scientific and Industrial Research

Energy generation from renewable sources such as solar and wind is becoming an attractive option, considering global and local trends towards sustainable planning and development. Energy planning requires a comprehensive understanding of the current and potential growth of various energy generating technologies, particularly the green technologies. In South Africa, there is a growing need to understand domestic solar usage.

An initial investigation was conducted to assess potential for detecting rooftop domestic solar installations (solar photovoltaic panels and solar water heaters) and potential growth monitoring using earth observation technologies. This investigation was tested on a small study site using various high-resolution images and looked promising. The estimated costs and data requirements, however, for acquisition of appropriate (very high spatial resolution) imagery were too high. As a result, a cost effective alternative methodology, which is presented in this paper, was investigated.

This methodology is based on integrating various sources of data to determine domestic solar usage with the aim of monitoring its growth over time. It includes segmenting households from a census 2011 sub-place data frame into units (clusters) with similar solar usage behaviour, sampling from such units to determine proportions of solar installations using existing sources of earth observation data, and extrapolating from the sample for the whole country.

Keywords: Clustering, sampling, integrated data approach

Coauthors: J Holloway (Council for Scientific and Industrial Research) and R Koen (Council for Scientific and Industrial Research)

Stream: Applied Statistics and General

A Statistical Analysis of the Presence of Form in Cricket

Christopher Dunderdale - Nelson Mandela University

Cricket is a sport with a significant amount of public interest, particularly in South Africa, and the existence of periodic good and bad form in cricketers has long been debated. The current study investigates and proposes some approaches for the detection of form in batsmen. A statistical evaluation of form in cricket batsmen by Durbach and Thiart (2007) found no basis for form in this context. However, the current study revisits this evaluation with a more complete data set. This comprehensive data set allows for a thorough assessment of the presence of form in batsmen using the runs test and survival analysis approaches of the original study. Furthermore, the current study proposes and assesses the use of alternative form detection methods in cricket.

Keywords: Cricket; Non-Parametric; Survival Analysis

Coauthors: W Brettenny (Nelson Mandela University) and C Clohessy (Nelson Mandela University)

Stream: Young Statisticians

A Statistical Assessment of Solar Resource Data Across Multiple Sites in South Africa

Kirstie Eastwood - Nelson Mandela University

Fossil fuels remain the primary source of energy accounting for about 78% of the world's total energy consumption. The high dependency on fossil fuels has created, among other concerns, an awareness of the need to reduce the environmental impact of carbon emissions. As a result, growth in the renewable energy sector has received extensive attention within the academic community. The primary beneficiaries of this growth have been engineers, economists and physicists. However, there is a need for expert statistical support as the research becomes more advanced. This research assesses the quality of solar resource data made publicly available by the Southern African Universities Radiometric Network, a collective group of

universities that provide high resolution, ground-based solar radiometric data.

Keywords: Fossil fuels, SAURAN, solar radiometric data

Coauthors: C Clohessy (Nelson Mandela University) and G Sharp (Nelson Mandela University).

Stream: Young Statisticians

A point process model for pulses of the DPT from an image

Inger Fabris-Rotelli - University of Pretoria

The Discrete Pulse Transform (DPT) is a multiscale decomposition of an image obtained using LULU smoothers. LULU smoothers are nonlinear and have many powerful advantages as such edge and total variation preservation. LULU smoothers have been developed for multiple dimensions – here we focus on images. A point pattern is extracted from an image using the DPT and modelled as a spatially clustered pattern, with trend in both the rows and columns of the image. The clustered nature arises naturally from the images structures which are well extracted by the DPT, and is confirmed prior to modelling. The fitted models perform well and extract the spatial intensity of the image. The DPT provides a representation of an image at all scales – from detail up to background shading - and enables informed analysis of the content and features of an image. This work expands the DPT image application into the spatial domain, a natural domain for image structures having obvious spatial dependence.

Keywords: Image processing, DPT, spatial statistics

Coauthors: A Stein (University of Pretoria and University of Twente)

Stream: Applied Statistics and General

Rank one noncentrality of complex Wishart type and its eigenvalue distributions

Johannes Ferreira - University of Pretoria

The eigenvalue distributions from a complex noncentral Wishart ($S=X^H X$) matrix have been the subject of interest in various real world applications, and have been studied where X is assumed to be a complex matrix variate normally distributed with nonzero mean M and covariance Σ . This paper focuses on a weighted analytical representation to alleviate the restriction of Gaussianity; thus allowing the choice of X to be complex matrix variate elliptically distributed and derives new results for eigenvalues of S , in this weighted representation, specifically for an underlying complex matrix variate t assumption. The distributions of the minimum eigenvalue of the complex Wishart type enjoy particular attention. This theoretical investigation has proposed impact in wireless communications systems, in particular the case where the noncentrality matrix has rank one which is of practical importance.

Keywords: Heavy tails, matrix variate, distribution theory, MIMO systems

Coauthors: Andriette Bekker (University of Pretoria)

Stream: Young Statisticians

Optimal mission duration for partially repairable systems operating in a random environment

Maxim Finkelstein - University of the Free State

A system failure during a mission can result in considerable penalties. In some cases it is cheaper to terminate the operation of a system than to attempt to complete its mission. This paper analyzes the optimal mission duration for systems that operate in a random environment modeled by a Poisson shock process and can be minimally repaired during a mission. Two independent sources of failures are considered, and for both cases the failures are classified as minor or terminal in accordance with the Brown-Proschan model. Under certain assumptions, an optimal time of mission termination is obtained. It is shown that, if for some reason a termination is not technically possible at this optimal time, the mission should be terminated within a specific time interval and, if this is not possible, it should not be terminated at all. Illustrative examples are presented. The influence of mission and system parameters on the mission termination interval is demonstrated.

Keywords: Premature mission termination; external shocks; expected profit; minimal repair; optimization

Coauthors:

Stream: Time Series and Stochastic Processes

Biplots based on Principal Surfaces

Raeesa Ganey - University of Witwatersrand and University of Cape Town

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. Biplots are considered as extensions of the ordinary scatterplot by providing for more than three variables. When the difference between data points is measured using a Euclidean-embeddable dissimilarity function, observations and the associated variables can be displayed on a non-linear biplot. A non-linear biplot is predictive if information on variables is added in such a way that it allows the values of the variables to be estimated for points in the biplot. Prediction trajectories, which tend to be non-linear are created on the biplot to allow information about variables to be estimated. The goal is to extend the idea of nonlinear biplot methodology onto principal surfaces. The emphasis will be on high dimensional data where the nonlinear biplot based on a principal surface will allow for visualization of samples and the predictive variable trajectories.

Keywords: Biplots; Principal surfaces; Nonparametric principal components; Multidimensional scaling

Coauthors: Sugnet Lubbe (Stellenbosch University)

Stream: Multivariate Data Analysis

Judicial decision-making: A statistical perspective

Nicola Gawler - University of Pretoria

Judicial decisions should be underpinned by fairness, bearing responsibility to the constitution and laws of a country or area. However, the objectivity of a judge's decision is often questioned by the public. The political affiliation and ideologies held by the judge are often thought to impact his/her application of the law. In this paper a statistical perspective on the fairness and impartiality of judicial decisions will be developed using a random forest as a prediction model for judicial decisions based on certain aspects of a case.

Keywords: Decision tree, Judicial impartiality, Random forest.

Coauthors: J van Niekerk (University of Pretoria)

Stream: Young Statisticians

Judicial decision-making: A statistical perspective

Nicola Gawler - University of Pretoria

Judicial decisions should be underpinned by fairness, bearing responsibility to the constitution and laws of a country or area. However, the objectivity of a judge's decision is often questioned by the public. The political affiliation and ideologies held by the judge are often thought to impact his/her application of the law. In this paper a statistical perspective on the fairness and impartiality of judicial decisions will be developed using a random forest as a prediction model for judicial decisions based on certain aspects of a case.

Keywords: Decision tree, Judicial impartiality, Random forest.

Coauthors: J van Niekerk (University of Pretoria)

Stream: Young Statisticians

Quantifying Eskom's Emissions Efficiency by Fitting The 3 Parameter Burr Distribution to the SO₂ Emissions in Electricity Power Generation using Coal in South Africa.

Elizabeth Girmay - University of the Free State

We quantify Eskom's emission efficiency by fitting a statistical distribution to Sulphur Dioxide (SO₂) monthly emissions (in kilogrammes per Gigawatt hour (kg/GWh) of energy produced and in milligrams per cubic Nano metre (mg/Nm³)), on Eskom's 13 coal fired power generating stations in South Africa. We aim to describe the emission of sulphur dioxide at Eskom's coal power stations using a statistical distribution. We propose the 3 parameter Burr distribution since it best fits the data of these 13 coal fired power stations. The

distribution fit allows one to quantify and manage the SO₂ emissions. Maximum likelihood is used to estimate the parameters, and various goodness of fit measures are employed. Further, objective Bayesian methods are employed to capture the uncertainty in the parameters, and to better estimate the probabilities of exceedances (emissions above a certain threshold) and extreme emissions.

Keywords: Emission, Eskom, Burr distribution, Goodness of fit, Sulphur dioxide, Bayesian Statistics

Coauthors: D Chikobvu (University of the Free State) and S van der Merwe (University of the Free State)

Stream: Applied Statistics and General

Spatiotemporal Mapping and Detection of Mortality Clusters Due to Cardiovascular Disease with Bayesian Hierarchical Framework Using Integrated Nested Laplace Approximation: Application in Kersa HDSS

Melkamu Dedefo Gishu - University of KwaZulu-Natal

Introduction: According to WHO's latest report cardiovascular disease (CVD) is the number one cause of death globally. Over three quarters of CVD deaths take place in low- and middle-income countries. Hence comprehensive information about the spatio-temporal distribution of mortality due to CVD is of interest.

Objectives: Detecting significant risk clusters of mortality within population based surveillance sites in eastern Ethiopia in Kersa HDSS.

Method: We fitted different spatio-temporal models within Bayesian hierarchical framework allowing different space time interaction for mortality mapping with integrated nested Laplace approximations to analyze mortality data which were extracted from the Kersa health and demographic surveillance system.

Results: The result shows non-parametric time trend models perform well. Mortality due to CVD increases during the study period, and administrative regions in the northern and south-east part of the study areas show a significant elevated risk.

Conclusions: The study has highlighted distinct clusters of mortality due to CVD both in space and in time within the study area. It can be viewed as a preliminary assessment step in prioritizing areas for more comprehensive research. Underlying contributing factors need to be identified and accurately quantified. Further research questions for more detailed investigation need to be raised.

Keywords: CVD, Bayesian hierarchical, spatio-temporal models, INLA, clusters

Coauthors: Henry Mwambi (University of KwaZulu-Natal), Nega Aseffa (Haramaya University) and Sileshi Fanta (University of KwaZulu-Natal)

Stream: Young Statisticians

Crocodile Movements from the Olifants Gorge – insights from Hidden Markov Models

Victoria Goodall - Nelson Mandela University

Hidden Markov models (HMMs) are commonly used to model animal movements. Terrestrial mammals make behavioural choices largely based on the distribution of resources. These are typically distributed in two dimensional space across a landscape with temporal variability. It is unclear how HMMs perform when resources are distributed linearly. During 2008/9 over 160 crocodile carcasses were found in the Olifants and Letaba rivers in the Kruger National Park. Post mortems on several animals revealed that the deaths were caused by a condition known as pansteatitis. This condition is caused by a depletion of the antioxidants in the body and results in hardening of the crocodile's fat reserves. This causes the animal to lose mobility and death results from starvation or drowning. Researchers fitted a number of live crocodiles with GPS tracking devices. The data are unique since the crocodiles are confined to these rivers and the movement is linear in an upstream or downstream direction only. We illustrate the use of the HMMs to uncover latent behavioural states which are then used to develop a behavioural profile for the tracked crocodiles. We investigate the influence of the direction of the movement in the modelling process and the ability of the models to uncover the behavioural patterns of the crocodiles. We use these to make inferences about potential mechanisms associated with behaviour that could compromise crocodile persistence when new diseases emerge in populations.

Keywords: Hidden Markov models, linear movement, crocodile

Coauthors: SM Ferreira (South African National Parks)

Stream: Applied Statistics and General

A pure jump interest rate model for the South African market.

Gerrit Grobler - North West University

In the field of Financial Mathematics stochastic differential equations are used to describe the dynamics of interest rates. An example is a model for the short rate, which is a mathematically defined rate not directly observable in any market. However, observable short dated rates such as the 91-day Treasury bill rate and the 3-month Johannesburg Interbank Agreement Rate (JIBAR) can be used as proxies of the short rate. With applications to price interest rate derivatives, diffusion processes are popular models for the short rate due to its analytical tractability. However, in the South African market no diffusion component is evident, specifically at low interest rate levels. This conclusion is made both through analysing historical interest rates as well as by testing for jumps. A pure jump model is fitted to the historical 3-month JIBAR. As a result a nonstationary compound Poisson process, with stably distributed jumps is suggested as a South African short rate model.

Keywords: Short rate, nonstationary compound Poisson process, diffusion process, jump diffusion process, pure jump process.

Coauthors:

Stream: Young Statisticians

A longitudinal COM-Poisson regression model with random effects

Freedom Gumedze - University of Cape Town

Longitudinal count responses can be modeled using the Poisson generalized linear mixed model to account for the correlations underlying the responses collected repeatedly over time. The count responses may be under-dispersed or over-dispersed relative to the Poisson distribution. The COM-Poisson regression model has been used to model independent count data which are under-dispersed or over-dispersed. Attempts to extend this model to the longitudinal case have focused on modelling the autocorrelation structure in the repeated count data explicitly. We consider the modelling of the general correlation structure of repeated count data using random effects. We use additional random effects to detect extra-variability in the data which is not accommodated by the COM-Poisson model. This extra-variability may be used to detect outliers at the observation-level or subject-level. We illustrate the methodology using a real data set.

Keywords: COM-Poisson longitudinal model, Generalized linear mixed effects model, Multivariate longitudinal data, Over-dispersion, Outlier detection, Random effects, Under-dispersion

Coauthors:

Stream: Biostatistics

Minimum information for training a classifier

Catherine Halsey - University of Pretoria

Classifier accuracy is important and can be improved by increasing the size of the sample data on which the classifier is based. However, in experimental cases it is not always possible to obtain enough of the required data to accurately train the classifier, as even very large data sets may not contain enough information, and access to the information may become computer intensive. For this reason a sequential method of training classifiers can be useful. This paper proposes a sequential method which seeks to sample the minimum number of observations necessary to train a classifier to estimate the feasible minimum rate of misclassification, the Bayes error, and ensure that the rates of misclassification are within this error. This method of classifier training gives the researcher more control over the process by specifying when the sequential procedure should be stopped. The method is not restricted to any single method of classification and it never seeks to obtain an unfeasibly low misclassification rate.

Keywords: Bayes error, Fixed-width confidence interval, Classifier training.

Coauthors: F Kanfer (University of Pretoria) and S Millard (University of Pretoria)

Stream: Young Statisticians

The application of natural language processing in the prediction of customer experience

Cindy Hayward - University of Stellenbosch

The data used in this study originates from a popular Mobile Network Operator's (MNO) Customer Care

Centre (CCC). Natural Language Processing (NLP) is used to analyse text messages received from customers in an attempt to gauge and classify the common problems they may have experienced with the company. In conjunction with using NLP techniques, this analysis makes use of two classifiers, namely Extreme Gradient Boosting (XGBoost) and Random Forests (RF), to classify text responses to commonly occurring Problem Groups (PGs). This study uses the XGBoost classifier as a benchmark classifier, and compares it to the results of a RF classifier with the goal of finding an model that outperforms the benchmark. This analysis aims to successfully utilise NLP techniques for cleaning and organising the text data. Secondly, this analysis aims to compare the effectiveness of an XGBoost classifier to a RF classifier, and further determine whether the RF classifier can outperform it.

Keywords: Natural Language Processing, Extreme Gradient Boosting, Random Forests

Coauthors: Wagenaar, B (Department of Statistics and Actuarial Science, Stellenbosch University)

Stream: Young Statisticians

On a Hazard Rate Processes with Delays after Shocks

Nil Kamal Hazra - University of the Free State

Distinct from conventional shock models when a failure of a system or accumulation of the corresponding damage occurs immediately after a shock, we consider a setting when these consequences appear with a random delay, which can happen in various applications. In our model, a shock acts directly upon the failure rate of a system. This shock, after a random time, can result either in a failure or in the increase in the failure rate by a random amount. We consider the corresponding hazard rate process and derive expressions for the survival probability and the failure rate of a system operating in a random environment modelled by the nonhomogeneous Poisson process of shocks with the delayed impacts. The asymptotic behaviour of the failure rate is studied in detail.

Keywords: Failure rate; hazard rate process; nonhomogeneous Poisson process; stochastic intensity

Coauthors: Maxim Finkelstein (University of the Free State) and Ji Hwan Cha (Ewha Womans University)

Stream: Time Series and Stochastic Processes

A regression analysis application to ecological drought vulnerability assessment: Eastern Free State, South Africa

Bernard Moeketsi Hlalele - University of the Free State

Implementation of adequate measures to assess and monitor drought events are major global challenges to water resources management. Consistent with the current droughts events in the globe, Free State Province in South Africa was one of the provinces declared struck by drought disaster in 2015. The objective of this study was to assess and monitor ecological drought vulnerability from historical data records drawn from Bethlehem airport station. Four ecological drought indicators over a 30 year period (1986-2016), were used to compute vulnerability time series and a composite ecological drought vulnerability index. A non-parametric homogeneity test was conducted prior to further analysis to avoid spurious results, where all data sets were homogeneous. The computation considered the functional relationship of drought and the selected indicators. All indicators' values were normalised to generate vulnerability time series. A normal probability distribution fitted well to the series using Kolmogorov-Smirnov test. For this reason, a Pearson correlation coefficient was used to check if any significant trends existed. The results revealed a significant annual trend in the series. Findings were concluded with a five year (2017-2021) forecast with increasing ecological drought vulnerability indices. The study therefore recommends that authorities put proactive relevant measures in place to protect the environmental and other ripple adverse effects on communities' livelihood.

Keywords: Ecological drought, vulnerability, regression analysis, disaster, composite vulnerability index

Coauthors:

Stream: Applied Statistics and General

Interactive Clustering using Projections

David Hofmeyr - Stellenbosch University

Model selection in clustering is challenging unless strict assumptions are placed on the generative process underlying the data, usually in the form of a parametric mixture model. In the absence of such assumptions,

and when data are so high-dimensional that fitting optimal parametric models becomes computationally prohibitive, low dimensional visualisations of the data which reveal their cluster structure can be invaluable in selecting an appropriate model. For more than a few clusters, however, it is not generally possible to provide a single visualisation which reveals the complete cluster structure. In this case it is preferable to adopt a hierarchical tree-like model in which a simple partitioning rule is applied at each node in the tree. In this way the individual partitions contributing to the model can be visualised, and thus validated, separately. This work uses a recently developed computationally efficient method for finding optimal hyperplane separators for clustering to provide an interactive recursive clustering algorithm. Each hyperplane generates a visualisation of the data (and the partition) by combining projections of the data onto the normal vector to the hyperplane and their principal components within the null space of this normal vector. Such partitions can be included in the model, discarded, or modified based on user specifications. Bootstrap and cross validation methods are provided to aid the user in making robust decisions.

Keywords: Clustering, Interactive, Projection, Projection Pursuit, Hyperplane, Normalised Cut

Coauthors:

Stream: Multivariate Data Analysis

Filtered DPOT and PORT models for high quantile forecasting

Chun-Kai Huang - University of Cape Town

The recently proposed duration-based peaks-over-threshold (DPOT) and peaks-over-random-threshold (PORT) methodologies are extended to amalgamate with various GARCH-type filters in high quantile estimation. This two-stage process allows for unconditional model specification of various stylised facts commonly inherent in financial data (for example, long memory property, leverage effects, and persistence), while at the same time merging the effects of DPOT and PORT in terms of model innovations. The performance of the proposed procedures is evaluated against classical extreme value models for financial risk forecasting of returns in various global indices.

Keywords: Extreme quantile; DPOT; PORT; GARCH; asset returns.

Coauthors: Chun-Sung Huang (University of Cape Town) and Ivivi Joseph Mwaniki (University of Nairobi)

Stream: Financial and Business Statistics

Forecasting Stock Market Volatility in the Presence of Structural Breaks: An Application to Value-at-Risk Estimation in South Africa

Chun-Sung Huang - University of Cape Town

This paper investigates the empirical evidence of structural breaks in stock return volatility for the South African Equity Market by using the Modified Iterative Cumulative Sum of Squares procedure. It then proposes an adjustment to the standard GARCH(1,1) and GJR-GARCH(1,1) models by incorporating endogenously identified structural breaks into the determination of the estimation window used to forecast volatility and explicitly imposing a lower bound on this window size. This is evaluated against a series of static GARCH models, including those designed to cater for the leverage and long memory effects inherent in financial time series. Our out-of-sample result shows that the structural break model outperforms all static GARCH models over the forecasting horizon. The evaluation criteria comprise a range of statistical and risk-management (Value-at-Risk) loss functions.

Keywords: ICSS algorithm, stock return volatility, structural breaks, GARCH, GJR-GARCH

Coauthors: L Reddy (University of Cape Town), D Pillai (University of Cape Town) and C-K Huang (University of Cape Town)

Stream: Financial and Business Statistics

Coverage probabilities and average lengths of fiducial generalized confidence intervals for model parameters and quantiles of the Generalized Extreme Value distribution

Peter Iiyambo - University of Namibia

Inference for the model parameters and quantiles of location-scale-shape (LSS) families of distributions is usually based on maximum likelihood, method of moments, probability-weighted-moments, or likelihood moments, among others. However, thus far there appears to be limited literature on generalized inference

for LSS families of distributions. We propose a method for construction of fiducial generalized confidence intervals (FGCIs) that involves the simulation of conditional fiducial generalized pivotal quantities (CFGPQs) using a Gibbs sampler. We study the coverage probabilities and average lengths of FGCIs for the model parameters and quantiles of the Generalized Extreme Value (GEV) distribution. Overall, simulation results show that the Gibbs sampler using rank-based CFGPQs produces FGCIs with generally good properties when the shape parameter is less than one. However, the Gibbs sampler algorithm using rank-based CFGPQs does not work well in the case of GEV for the shape parameter greater than one.

Keywords: Coverage probability, fiducial generalized confidence interval, Generalized Extreme Value distribution, location-scale-shape family, rank-based fiducial method, conditional fiducial generalized pivotal quantity, Gibbs sampler

Coauthors: R Schall (University of the Free State)

Stream: Applied Statistics and General

A Skew-t-normal Generalised Autoregressive Score (STN-GAS) Model

Stefan Janse van Rensburg - Nelson Mandela University

A generalised autoregressive score (GAS) model based on a skew-t-normal (STN) distribution is proposed. The resulting STN-GAS model is capable of modelling both conditional volatility and conditional skewness. Estimation and inference are discussed. A simple application to equities listed on the Johannesburg Stock Exchange is presented.

Keywords: Asymmetry; generalised autoregressive score model; skew-t-normal distribution; conditional volatility

Coauthors: GD Sharp (Nelson Mandela University)

Stream: Young Statisticians

The use of Structural Equation Models to explore direct and indirect paths

Esme Jordaan - South African Medical Research Council

Structural equation modeling (SEM) is a series of statistical methods that allows for complex relationships between one or more latent independent variables and one or more latent independent variables. SEM is a powerful analytical tool to examine complex causal relationships because it is superior over other correlational methods such as regression, as multiple variables that are analysed simultaneously, and latent factors, reduce measurement error. SEM allows one to test a hypothesized path model and to identify significant paths as well as testing of mediation through direct and indirect path relationships between the risk factors and the outcome. In the example, the relationships between exercise associated muscle cramping (EAMC) and risk factors were explored. The objective of the analysis was to explore complex relationships among the risk factors associated with EAMC using knowledge about the associations of the risk factors with EAMC. We hypothesized a structural mediation path model with direct and indirect paths of interest that we wished to explore.

Keywords: Structural equation model, mediation, paths

Coauthors: M Schwellnus (University of Pretoria)

Stream: Biostatistics

Detecting influential communities in the modelling of clustered child survival data in Malawi

Tsirizani Kaombe - University of Malawi

A number of studies investigating child survival probabilities using complex survey data are now routinely done through multilevel survival models implemented in some statistical software. However, little attention is being paid to assessing influence of outlying clustering units, whose inclusion or exclusion might affect both parameter and variance estimates in the model. We endeavoured to assess the influence of some groups of Malawian children when a survival model is fitted to national data whose sampling units were enumeration areas (EAs). The model fitting was done both for the complete data set and for a sub-set formed by eliminating some of the EAs, while observing how this affects various measures such as parameter estimates.

Keywords: Clustered data, survival model, influence measure

Coauthors: S Manda (South African Medical Research Council) and S White (Liverpool School of Tropical Medicine)

Stream: Biostatistics

Providing insight on the WHO's HIV indicators for the population aged 15 – 24 years using the interval censoring time to event analysis: An empirical study using the 2012 South African National HIV Prevalence, HIV Incidence, Behaviour and Communication Survey

Innocent Karangwa - Stellenbosch University

According to the latest report by the Human Science Research Council (HSRC), HIV prevalence has increased over the years. This increase is partly due to the huge increase in antiretroviral programmes that save many lives in South Africa. The prevalence varies by the gender and race of the respondents. The United Nations General Assembly Special Session on HIV/AIDS (UNGASS) has suggested several indicators for monitoring the HIV epidemic at a country level. Some of these indicators are specific to the age group 15-24. The HSRC SABSSM surveys also focus on this age group and have reported some of the UNGASS indicators in their 2012 survey report. Using parametric / non-parametric survival models for interval-censored and time to event data, this study seeks to determine whether or not the estimates of HIV indicators provided by the 2012 report are in line with our estimates, and to expand and enhance the information available in the 2012 survey regarding HIV indicators in the population aged 15-24 years. Our method involves using the reported age of sexual debut and the age at the time of the survey as the interval-censored time to event for HIV positive 15-24 year old participants to reflect the true state of information on the timing of the HIV infection given the cross-sectional nature of the survey. For HIV negative participants, the age at the time of the survey is the right censoring indicator.

Keywords: HIV, interval censoring, WHO indicators

Coauthors: C Lombard (South African Medical Research Council and Stellenbosch University)

Stream: Biostatistics

Identifying Risk Factors to the Survival Time of Adult HIV Patients on Antiretroviral Therapy at Windhoek Public Hospitals, Namibia. A Retrospective Study with Proportional Hazard Model.

Simon Kashihalwa - University of Namibia

In Namibia free ART service has been expanding but there is limited information regarding treatment outcome and risk factors for patient survival. The objective of this study was to identify risk factors to the survival time of HIV patients (15 years and older) on ART, using data obtained from Windhoek public Hospitals. The K-M Method was employed to estimate the survival probabilities and log-rank tests were used to determine differences in survival between groups. The PH Regression Method was then used to identify risk factors for all-cause mortality. This study used 450 HIV/AIDS patients (15 years and older) who initiated treatment between 2008 and 2010 and were followed until the end of 2014. Of these 59% were male and 41% were female. The accounted mortality and censored patients in the study period were 24% and 76%, respectively. The estimated median survival time was 1328 days. Survival time of the HIV patients was significantly related to Age at HAART, CD4 count, WHO Clinical stage and Sex. In the absence of ART, it is well known that the quality of life and the lifespan of a patient with HIV/AIDS are markedly reduced.

Keywords: HIV/AIDS, Survival Analysis, Risk Factors, ART

Coauthors: I Maposa (NUST)

Stream: Biostatistics

Latent trait analysis of the effects of categorical covariates on binary outcomes using structural equation modeling

Robert Keli - University of Kwazulu-Natal

Causal modeling has found wide application in medical research. It seeks to establish the presence and strength of relationships that exist between the predictor variables and the outcome(s). Such effects can either be direct or through (observed and / or latent) mediators. Two models, one with observed and the other with latent mediators are developed and analyzed in both frequentist and Bayesian paradigms using the structural equation modeling technique. Based on empirical data from the Kenya Aids Indicator Survey of 2007, the effects of the categorical covariates on the HIV status of Kenyan women aged between 15 - 64 years are estimated. The developed models are implemented in Mplus and the maximum likelihood,

bootstrap and Bayes parameter estimates are presented. The results show that the Bayes estimates are more efficient since they exhibited the lowest variance. The latent mediator model had the lowest Bayesian information criterion (BIC) fit index measure, hence the best. Also, the effects of the risk factors were found to be consistent with results from previous studies on HIV/AIDS.

Keywords: Structural Equation Modeling; Categorical data; HIV/AIDS; Mediation analysis; Maximum Likelihood; Bootstrap estimation, Bayes estimation

Coauthors: Henry G Mwambi (University of Kwazulu-Natal) and Elphas L Okango (University of Kwazulu-Natal)

Stream: Biostatistics

SAS tools and SAS & Open source tools for the 3 things

Paul Kent - SAS

Building Great models is not enough. Your models won't be that great if you don't have a good data strategy for raw material, and you won't get good returns if you don't find good ways to use your models quickly. The talk explores how data warehousing has changed in this age of analytics, and uses examples from business around the globe. At the plenary we'll be talking about 3 things you need to do well to get value from analytics. The workshop will feature three segments on those same 3 themes: 1. Tools for maintaining your data lake, 2. Tools for building models and forecasts, 3. Tools for deploying models

Keywords:

Coauthors:

Stream: Workshop

3 things you need to do well to get full value from SAS analytics and SAS & Open source analytics

Paul Kent - SAS

Building Great models is not enough. Your models won't be that great if you don't have a good data strategy for raw material, and you won't get good returns if you don't find good ways to use your models quickly. The talk explores how data warehousing has changed in this age of analytics, and uses examples from business around the globe.

Keywords:

Coauthors:

Stream: Plenary

Data Lakes - is this really any different than the data warehouse we were building at the turn of the century.

Paul Kent - SAS

What makes a modern Data Lake strategy different from the Enterprise Data Warehouse initiatives from days gone by. Can they be more "just-in-time", more agile, and more flexible? The talk dives into how data strategy has evolved, and what technological advances including SAS & Open Source have led to changed thinking.

Keywords:

Coauthors:

Stream: Business

Extension and alternative to the alpha-mu fading model

Michaela Laidlaw - University of Pretoria

This paper introduces the alpha-mu fading model within the elliptical class. The distribution's characteristics are discussed and its feasibility as a fading model in wireless communications systems is investigated.

Keywords: Average bit-error rate, elliptical class, fading model, generalized gamma, outage probability, signal-to-noise ratio

Coauthors: A Bekker (University of Pretoria) and JT Ferreira (University of Pretoria)

Stream: Theoretical Statistics

Sparse linear discriminant analysis for high-dimensional data: Techniques and packages in R

Morne Lamont - Stellenbosch University

Techniques for the classification of observations into known classes have many practical applications. Statistical classification is also a very active research area for many Statisticians. The most well-known classification (discriminant) technique was proposed by Fisher (1936) and is called Fisher's linear discriminant analysis. Fisher's LDA is a powerful classification tool, but suffers from a singularity problem when $n \ll p$. Over the past few decades, many researchers have made proposals to solve this singularity problem. Many of these proposals also perform variable selection by using the LASSO and elastic net penalties. This paper is a brief summary of some of these proposals, often referred to as sparse discriminant analysis techniques. The paper highlights the R packages and functions available to users.

Keywords: Sparse discriminant analysis; Penalized LDA; Sparse partial least squares; Shrunken centroids regularized discriminant analysis

Coauthors:

Stream: Multivariate Data Analysis

Recent approaches to analysing nutrition data

Ria Laubscher - South African Medical Research Council

Traditionally, nutrition data are examined by analysing the nutrients and/or foods, but people do not eat nutrients, they eat meals consisting of different foods with complex combinations of nutrients. This single-nutrient approach does not account for the complicated interactions among nutrients of free-living individuals. Furthermore, the high level of correlation between some nutrients complicates the examination of their separate effects. Nutrient intakes are associated with certain eating patterns and therefore single-nutrient analysis might be confounded by the effect of these eating patterns. In recent years, several authors have proposed that one must rather study the overall eating patterns by considering how foods and nutrients are consumed in combination. This will overcome the abovementioned limitations of nutritional analyses. By analysing eating patterns, one can use the collinearity of nutrients and foods to advantage, because the patterns are characterized based on habitual food consumption (Food Frequency Questionnaire). Therefore, these patterns will be closer to real-world situations in which eating patterns consist of nutrients that occur together in common foods. This proposed method will be illustrated using dietary data obtain by administering a FFQ to South Africans in the North-West province. The data reduction technique most suited for the analyses will be discussed.

Keywords: Nutrition, Data reduction

Coauthors: E Wentzel-Viljoen (North-West University) and E Vorster (North-West University)

Stream: Biostatistics

Biplot visualisations of interactions in a bilinear model having a qualitative dependent variable

Niel Le Roux - Stellenbosch University

Conventionally, a bilinear model of a quantitative dependent variable classified by a row and a column variable is fitted by first fitting the main effects followed by fitting a biadditive model to the residual table. This gives the ordinary least squares optimal fit and the residual table can be used as input to a biplot for visualising the interactions in the table. In practice two issues are to be addressed: first, whether a dependent variable is categorical; second the optimal approximation of the interactions. We extend the proposal by Fisher (1938) to obtain optimal quantifications for a qualitative dependent variable by maximising the row and column sum-of-squares relative to the residual sum-of-squares. We show that with a categorical dependent variable, it is beneficial to compute quantifications by maximising the sum-of-squares of the biadditive part of the model relative to the residual sum-of-squares. This approach leads to an optimal biplot display of interactions in a two-way table. Other ways of partitioning the contributions from main effects and biadditivity are also discussed. Writing the original short form of the data i.e. the two-way table, in long form as a subjects \times variables data matrix opens up other ways of constructing biplots. The various extensions enable a user to choose not only whether to optimally approximate the main effects or the interaction effects but also to consider several different biplots for exploring the data visually.

Keywords: Biadditive models, biplots, categorical data, modelling, multiplicative interactions, non-additive

interactions, optimal scores

Coauthors: S Lubbe (Stellenbosch University) and J Gower (The Open University)

Stream: Multivariate Data Analysis

Short term load forecasting using quantile regression with an application to the unit commitment problem

Moshoko Emily Lebotsa - University of Venda

Short term probabilistic load forecasting is essential for any power generating utility. We develop a short term load forecasting model for peak demand hours (i.e. from 18:00 to 20:00) using South African electricity demand data for the years 2010 to 2013. Quantile regression is proposed together with a semi-parametric additive model to estimate the relationship between the demand and the driving factors. A semi-parametric additive model is considered because of its ability to capture the nonlinear relationship between load demand and temperature. Additionally, Lagrange relaxation and mixed integer programming techniques are used on the forecasts obtained in order to find an optimal number of units to commit (switch on or off). A feasible solution to the unit commitment will help utilities meet the demand with minimal costs. This modelling framework is useful to power utility companies for the scheduling and dispatching of electrical energy at the minimal possible cost.

Keywords: Lagrange relaxation, LASSO, quantile regression, mixed integer programming, short term peak load forecasting, unit commitment

Coauthors: C Sigauke (University of Venda), A Bere (University of Venda), R Fildes (University of Lancaster), J Boylan (University of Lancaster) and L Nedzingahe (Eskom)

Stream: Young Statisticians

The analysis of interval-censored observations

Emmanuel Lesaffre - I-Biostat, School of Public Health, KU Leuven, Leuven, Belgium

Interval-censored data occur in basically all empirical research, but is rarely taken into account properly. Appropriate methodology has been developed for a large number of problems. Despite the available statistical technology, interval-censoring is often swept under the carpet. A notorious example is the analysis of cancer trials with progression-free survival as outcome, in which almost invariably interval-censoring is ignored. We argue that one of the reasons that the appropriate technology is not used is the ignorance of many statisticians what the effects are of ignoring interval censoring, but also what software is available. In the forthcoming CRC Press book (Survival Analysis with Interval-Censored Data: A Practical Approach with examples in R, SAS and BUGS by Kris Bogaerts, Arnošt Komárek, Emmanuel Lesaffre, 2017), we give an overview of the pitfalls of ignoring interval censoring, of the different approaches both in a frequentist and a Bayesian context and the currently available statistical software.

Keywords:

Coauthors: Kris Bogaerts, Arnošt Komárek

Stream: Workshop

Why interval censoring should not be ignored

Emmanuel Lesaffre - I-Biostat, School of Public Health, KU Leuven, Leuven, Belgium

We consider here methods to analyze interval-censored survival times. Interval censoring occurs when it is only known that the event happened inbetween two examinations. Well-known examples of an interval-censored time are the time until HIV, AIDS, the emergence of a tooth, etc. Most often interval censoring is not appropriately addressed in a statistical analysis and dealt with by methods that handle right censoring of data, e.g. by replacing the interval by the mid-point. Despite several published results it is still too often believed that ignoring the interval-censored character of the data has a minimal impact on the results and conclusions of the statistical analysis. In this contribution we summarize the literature on interval censoring largely from a practical point of view under the frequentist and a Bayesian paradigm. It will be also discussed when it is important to take interval censoring into account.

Keywords:

Coauthors:

Stream: Plenary

Assessing Variance Components of Multilevel Models for Social Science Data: Application to Teenage Pregnancy Data

Marothi Peter Letsoalo - University of Limpopo

Most social data are longitudinal and additionally multilevel in nature, which means response data are grouped by attributes of some cluster. This work intended to explore and fit teenage pregnancy (TP) census data gathered from 2011 to 2015 by the Africa Centre at Kwa-Zulu Natal in South Africa. The exploration of this data revealed a two level pure hierarchy data structure of TP status for some year/s that are nested within female teenagers. Model building of this work, first, fitted a logit generalised linear model (GLM) under the assumption that TP measurements are independent between females and secondly, a MLM of female random effect. The effect that census year and other three females characteristics has on teenage pregnancy was examined. A better fit MLM indicated, for an additional year, a 0.203 decrease on the log odds of TP while GLM suggested a 0.21 decrease and 0.557 increase for each additional year on age and year, respectively. A GLM with only year effect uncovered a fixed estimate that is higher, by 0.04, than that of a better fit MLM. The inconsistency in the effect of year was caused by a significant female cluster variance of approximately 0.35. The VCs revealed that 9.5% of the differences in TP lies between females while 0.095 similarities (scale from 0 to 1) are for the same female. Furthermore, it was also revealed that year does not vary within females.

Keywords: Multilevel Model, Generalised Linear Mixed Model, Variance Components, Hierarchical Data Structure, Social Science Data, Teenage Pregnancy

Coauthors: Christel Faes (Hasselt University), Yehenew G Kifle (University of Limpopo) and Lesaoana Maseka (University of Limpopo)

Stream: Applied Statistics and General

Approximations for the distribution of the product of independent beta random variables

Theodor Loots - University of Pretoria

Various general procedures exist for deriving representations of the exact distribution of the product of independent random variables. Such products are related to the distribution of many LRT statistics used in testing, for example, the equality of covariance matrices (equal sample size case), sphericity, independence of several groups of variables, compound symmetry and circularity. Here, specifically, approximations to the distribution of the products of independent beta random variables are presented, and their performance compared to the corresponding exact representations.

Keywords: Computational methods, distribution theory, hypothesis testing, inference

Coauthors: A Bekker (University of Pretoria) and FJ Marques (Universidade Nova de Lisboa)

Stream: Theoretical Statistics

clusboot: An R package for visualising bootstrapping of cluster analysis

Sugnet Lubbe - Stellenbosch University

In the literature the bootstrap is used in a cluster analysis context to assist in determining the number of clusters, bootstrapping the similarity measure at a merger, or the sum of squared distances for the clustering solution. All these measures are single scalar quantities associated with the cluster analysis. Here, the focus is not on bootstrapping a single quantity, but replicating the complete clustering solution a large number of times. Any clustering algorithm can be plugged into the bootstrap analysis. The package clusboot provides a visual summary of the bootstrap results with multidimensional scaling. Furthermore, a cluster silhouette plot is constructed to represent the stability of the original clustering solution. The package will be illustrated with well-known example data sets as well as a practical application in a psychiatry context.

Keywords: Bootstrap, Cluster analysis, Multidimensional scaling

Coauthors: S Lubbe (Stellenbosch University)

Stream: Multivariate Data Analysis

Customer churn and fraud prediction in SPSS Modeler

Michelle MacDevette - OLSPS Analytics

OLSPS Analytics creates business solutions using predictive analytics across a range of industries. With a focus on machine learning OLSPS has taken the lead in predictive analytics consulting in sub-Saharan Africa. The solution offerings are implemented using software tools such as R, SAS, Python, IBM SPSS Statistics, and SPSS Modeler. We will show how to take advantage of SPSS Modeler to rapidly implement machine learning solutions to real world problems. In particular, a demonstration of a customer churn and fraud prediction solution will highlight the value of predictive modeling.

Keywords: Churn, Fraud prediction, IBM, SPSS Modeler, Predictive analytics

Coauthors: Rikus Combrinck (OLSPS Analytics)

Stream: Financial and Business Statistics

A time-series model for underdispersed or overdispersed count

Iain MacDonald - University of Cape Town

Time series of unbounded counts (that is, nonnegative integers) commonly display overdispersion relative to the Poisson. Such a series can be modelled by a hidden Markov model with Poisson state-dependent distributions (a Poisson-HMM), since a Poisson-HMM allows for both overdispersion and serial dependence. Time series of underdispersed counts are less common, but more awkward to model; a Poisson-HMM cannot cope with the underdispersion. However, if in a Poisson-HMM one replaces the Poisson distributions by Conway-Maxwell-Poisson distributions, one obtains a class of models which can allow for both under- or overdispersion (and serial dependence). In addition, this class can cope with the combination of slight overdispersion and substantial serial dependence, a combination that is difficult for a Poisson-HMM to represent. We discuss the properties of this class of models, and use direct numerical maximization of likelihood to fit a range of models to three published series of counts which display underdispersion, and to a series which displays slight overdispersion plus substantial serial dependence.

Keywords: Time series; counts; underdispersion; overdispersion

Coauthors:

Stream: Time Series and Stochastic Processes

Inference on the Destructive COM-Poisson Gamma Cure Rate Model

Jacob Majakwara - University of the Witwatersrand

We consider a model that assumes the occurrence of the event of interest to undergo a destructive process of the initial risk factor, and what is recorded is the undamaged portion of the original number of risk factors, which provides a realistic interpretation of the biological mechanism of the event of interest. By assuming a COM-Poisson distribution for the initial risk factors and gamma distribution for lifetime, the steps of an EM algorithm are developed to calculate the MLEs of the model parameters. An extensive simulation study is carried out to demonstrate the performance of the proposed estimation method, and finally a melanoma data set is analysed for illustrative purposes

Keywords: COM-Poisson, Competing cause scenario, EM algorithm, gamma distribution

Coauthors: S Pal (University of Texas at Arlington)

Stream: Biostatistics

Modelling the volatility of international tourist arrivals in Zimbabwe using a GARCH process.

Tendai Makoni - Great Zimbabwe University

In developing countries like Zimbabwe, tourism is a major economic driver. Modelling tourism demand and volatility is important, and this is the aim of this paper. SARIMA and GARCH models have better forecasting power, and better capture seasonality and volatility, than the simple regression model being used by the Zimbabwe Tourism Authority (ZTA) which may lead to spurious regression and forecasting failure according to Song et al. (2008). The logarithm of the monthly international tourist arrivals for the year 2000 to 2016 obtained from the ZTA is used. A time series plot indicated a stationary series, though with volatility clusters. A SARIMA(1,0,0)(1,0,1)₁₂ model fits the data well and indicated a future increase in the international tourist

arrivals. Accommodation and transport facilities must be well established, be of good quality and be readily available for the observed increase. The presence of ARCH effect on the SARIMA(1,0,0)(1,0,1)₁₂ model residuals resulted in fitting a GARCH(1,1) process under a GED distribution. This indicated future uncertainty in international tourist arrivals. Suitable actions and policies must be readily available to deal with future uncertainty. The ZTA should continue to market and improve tourist attraction centres in the country to both new and old visitors as this will create foreign currency earnings.

Keywords: SARIMA, ARCH, GARCH model, volatility.

Coauthors: D Chikobvu (University of the Free State)

Stream: Time Series and Stochastic Processes

Design of a side-sensitive double sampling \bar{X} control chart for monitoring the location process parameter

Jean-Claude Malela Majika - University of South Africa

This paper develops a new double sampling (DS) monitoring scheme, namely the side-sensitive DS \bar{X} chart, to monitor the process mean. We first give the operational procedure and thereafter the exact form of the probability of the in-control process and the average sample size for the proposed chart under the assumption of known process parameters. Finally, we investigate the performance of the new scheme by minimising the out-of-control average run-length and extra quadratic loss function. It was observed that the proposed chart presented a better overall performance than the existing DS \bar{X} charts and any other competing chart considered in this paper. An illustrative example is given to facilitate the design and implementation of the proposed chart.

Keywords: Process monitoring; DS scheme; side-sensitive DS scheme; performance measures; average sample size; average number of observations to signal; extra quadratic loss function

Coauthors: EM Rapoo (University of South Africa) and MA Graham (University of Pretoria)

Stream: Applied Statistics and General

Generating Functions in Branching Processes and Birth and Death Processes

Raeesa Manjoo-Docrat - University of the Witwatersrand

Birth and death processes and branching processes are stochastic processes that can be applied to fields and disciplines such as biology, economics and engineering. In general, the distributions and moments of these processes are difficult to obtain in explicit form. The use of generating functions makes computation much easier. We look at the theory, application and innovation of the use of generating functions in the analysis of birth and death processes and branching processes.

Keywords: Birth and Death Processes, Branching Processes, Generating Functions

Coauthors: F Beichelt (University of Witwatersrand)

Stream: Time Series and Stochastic Processes

Modelling nonstationary extremes in the lower Limpopo River basin of Mozambique

Daniel Maposa - University of Limpopo

We fit a time-dependent generalised extreme value (GEV) distribution to annual maximum flood heights at three sites: Chokwe, Sicacate and Combomune in the lower Limpopo River basin of Mozambique. A GEV distribution is fitted to six annual maximum time series models at each site, namely: annual daily maximum (AM1), annual 2-day maximum (AM2), annual 5-day maximum (AM5), annual 7-day maximum (AM7), annual 10-day maximum (AM10) and annual 30-day maximum (AM30). Nonstationary time-dependent GEV models with a linear trend in location and scale parameters are considered. The results show lack of sufficient evidence to indicate a linear trend in the location parameter at all the three sites. On the other hand, the findings reveal strong evidence of the existence of a linear trend in the scale parameter at Combomune and Sicacate, while the scale parameter had no significant linear trend at Chokwe. Further investigation also reveals that the location parameter at Sicacate can be modelled by a nonlinear quadratic trend; however, the complexity of the overall model is not worthwhile in fit over a time-homogeneous model. This study shows the importance of extending the time-homogeneous GEV model to incorporate climate change factors such as trend in the lower Limpopo River basin, particularly in this era of global warming and a changing climate.

Keywords: Nonstationary extremes, annual maxima, lower Limpopo River, generalised extreme value.

Coauthors: JJ Cochran (University of Alabama) and M Lesaoana (University of Limpopo)

Stream: Time Series and Stochastic Processes

Convolutional Neural Networks for Multi-Label Classification of Satellite Images

Jan Marais - Stellenbosch University

Deforestation has devastating effects on the environment, contributing to climate change and habitat loss, amongst others. Information regarding the location and cause of deforestation may assist authorities in improving their understanding of the problem. Therefore, a classifier which is able to label satellite images of a rainforest with various categories of land cover and land use, may prove very helpful. We pose this problem as a multi-label image classification problem, where each satellite image can potentially be annotated with more than one label. Convolutional Neural Networks (CNNs) have consistently proven their superiority in single-label image classification problems. However, it is unclear how to best extend CNNs to effectively deal with the unique challenges that arise in the multi-label setting. This work critically reviews proposals in the literature and compares them on a dataset representing more than 100000 labelled satellite images of the Amazon rainforest. We found that (1) training the network with the label-wise binary cross-entropy loss function is sufficient in dealing with label imbalance; (2) predictions from multi-scale feature maps facilitate the detection of smaller objects; and (3) explicitly modelling channel-wise dependencies of feature maps assists the network to exploit semantic and spatial label relationships.

Keywords: Multi-label classification, convolutional neural networks, image classification, remote sensing

Coauthors: S Bierman (Stellenbosch University)

Stream: Applied Statistics and General

Complex survey sampling methodology applied to the South African Census 2011 small area layer data

Thanyani Maremba - Statistics South Africa

Complex survey sample design (CSSD) is a probability sample developed using sampling procedures such as stratification, clustering, segmentation, probability proportional to size selection methods and weighting. CSSD intends to improve statistical efficiency, reduce costs and improve precision for sub-group analyses relative to a simple random sample. CSSD is one of the most challenging fields when applying statistical methodologies, including, but not limited to: non-satisfactory sample sizes, incorporation of the auxiliary information available on many levels and simultaneous estimation of characteristics in various sub-populations. In light of the prevailing challenges, this paper uses the theory of survey sampling and practical implementation of the CSSD methods employing census attributes. The methodology applies explicit stratification using natural area clusters; implicit stratification based on census attributes; and square-root allocation of provincial sample. As a result, the sample of the smallest province in South Africa, the Northern Cape was augmented. The first stage sampling units, i.e. small area layers (SALs), were selected using randomised PPS while the second stage units, i.e. households within SALs, were selected using systematic random sampling. A sample of 1,849 SALs was selected from 75,314 eligible SALs, and 17,076 households were selected from a sample frame of 524,308 simulated households.

Keywords: Sampling, Probability proportional to size, Stratification, Clustering

Coauthors: P Nyamugure and M Lesaoana

Stream: Official Statistics

Penalized bias reduction in extreme value estimation for censored Pareto-type data, and long-tailed insurance applications

Gaonyalelwe Maribe - University of Free State

Tail estimation using randomly censored data from a heavy tailed distribution receives growing attention, motivated by applications, for instance, in actuarial statistics. The bias of available estimators of the extreme value index can be substantial and depends strongly on the amount of censoring. We review the available estimators, propose a new bias reduced estimator, and show how shrinkage estimation can help to keep the MSE under control. A bootstrap algorithm is proposed to construct confidence intervals. We compare these proposals with the existing estimators through simulation. We conclude with a study of a long-tailed car insurance portfolio, which typically exhibits heavy censoring.

Keywords: Extreme value index; Pareto-type; Tail estimation; Random censoring; Bias reduction.

Coauthors: Jan Beirlant (KU Leuven and University of Free State) and Andréhette Verster (University of Free State)

Stream: Young Statisticians

A Quantitative Assessment of Municipal Revenue Sources in the Eastern Cape

Hellen Maribe - Statistics South Africa

Municipalities play a significant role in the social and economic development of society by ensuring that all people are provided with essential basic public services such as water, electricity, refuse removal, road maintenance, sewerage and sanitation. In South Africa municipalities are largely self-financed and depend on two main revenue sources: municipal own-revenue and intergovernmental transfers/government grants and subsidies (GGS). In the Eastern Cape inefficient revenue collection, particularly of property rates and service charges, which is mainly due to non-payment by consumers, presents a great financial risk to the municipality and impacts negatively on its fiscal capacity. This study presents an assessment of municipal revenue sources (service charges, property rates and GGS) and analyses the dependency of municipalities on GGS. The main data source for the study is the Municipal Annual Financial Statements (AFS) and Stats SA's financial census of municipalities (FCM). Using a quantitative approach the study found that municipalities in Eastern Cape are not fully efficient and effective in their revenue collection. The results also show a consistent increase in allocations of GGS to municipalities in Eastern Cape from 2010 to 2015.

Keywords: Non-payment, service charges, property rates, government grants and subsidies, municipality, Eastern Cape

Coauthors: Patrick Naidoo

Stream: Official Statistics

Risk management with long memory GARCH models: Empirical evidence from the USD/ZAR exchange rate

Harriet K Marima - University of KwaZulu-Natal

We evaluate Value-at-Risk (VaR) and Expected Shortfall (ES) for the USD/ZAR foreign exchange market. We adopt the FIGARCH, HYGARCH and the FIAPARCH models, models that have been proven to capture the long range dependence and volatility clustering nature of financial data very well, while the Generalized Error Distribution (GED), Student-t and Skewed Student-t distributions are adopted to capture the heavy tail and asymmetric behaviour of exchange rate series. The Anderson-Darling test is used to check for model adequacy, while the Kupiec Likelihood Ratio and the Dynamic Quantile tests are used to objectively compare performances of the three models in VaR and ES forecasting. Our findings confirm that the FIGARCH, HYGARCH and FIAPARCH models with GED and Skewed Student-t innovations are suitable for modelling USD/ZAR exchange rate returns and can be used in improving risk management assessment in the foreign exchange rate market, with the FIAPARCH model outdoing the other two models.

Keywords: Long memory, GARCH, Value-at-Risk, Expected shortfall, Heavy-tailed distributions

Coauthors: Knowledge Chinhamu and Retius Chifurira

Stream: Financial and Business Statistics

Overall Performance of Grade 12 Learners in the Eastern Cape and Mpumalanga Province

Pauline Masemola - UMALUSI

The collection of clustered observations is increasingly important in education. However, these data are often inappropriately analysed such that the effect of clustering is ignored. This study was aimed at comparing parameter estimates generated by naïve pooling and hierarchical models. The data were sourced from the Department of Basic Education through Umalusi council. The results of this study showed that naïve pooling than hierarchical linear model produced smaller standard errors. Also, naïve pooling gave falsely narrower confidence intervals. As a result, naïve pooling produced inefficient coefficient estimates. Therefore, the analyses of clustered data with binary end-points are best performed using statistical techniques that account for the clustering effect in situations where data are correlated, clustered or grouped.

Keywords: Clustered data, hierarchical model, naïve pooling, parameter estimates, confidence interval, standard error

Coauthors: ME Letsoalo (Tshwane University of Technology) and MA Lesaoana (University of Limpopo)

Stream: Statistics in Education

A Double Generally Weighted Moving Average Chart for Time Between Events

Hossein Masoumi Karakani - Department of Statistics, University of Pretoria

The concept of a control chart has been developed in the 1920s by Dr Walter Shewhart. Due to the rapid development of technology and increasing effort on process monitoring which led to high-quality processes, Shewhart-type attributes control charts are inefficient in detecting small changes for nonconformities. To overcome this shortcoming, an alternative approach is to use time-weighted control chart (also known as memory-based control chart) to monitor the time between events (TBE); these time-weighted control charts use all the information from the start until the most recent sample/observation to decide if a process is in-control or out-of-control. To this end, a generalized type of time-weighted control chart to monitor the TBE is proposed. This chart is called the Double Generally Weighted Moving Average Chart Time Between Events (DGWMA-TBE), which includes many of the well-known existing time-weighted control charts as special or limiting cases. An extensive simulation study reveals that the proposed DGWMA-TBE chart outperforms the Generally Weighted Moving Average (GWMA), Exponentially Weighted Moving Average (EWMA) and Shewhart charts at detecting small to moderate shifts.

Keywords: Average run length; DEWMA chart; DGWMA chart; Exponential; Time between events.

Coauthors: Human, S.W. (Department of Statistics, University of Pretoria) and van Niekerk, J. (Department of Statistics, University of Pretoria)

Stream: Applied Statistics and General

Peak electricity demand forecasting using partially linear additive quantile regression models

Norman Maswanganyi - University of Limpopo

The paper presents an application of partially linear additive quantile regression models in forecasting peak electricity demand using South African data from January 2007 to December 2013. Variable selection is done using the least absolute shrinkage and selection operator (Lasso) via hierarchical pairwise interactions in which the main effects (lower order variables) must be in the model if higher order interactions are included. One of the contributions of this paper is the inclusion of a nonlinear trend as one of the covariates which is determined using a penalized cubic smoothing spline.

Keywords: Additive models, Forecast combination, Lasso, Peak demand forecasting, Quantile regression.

Coauthors: C Sigauke (University of Venda) and E Ranganai (University of South Africa)

Stream: Applied Statistics and General

Forecasting capacity loss in Eskom generation

Akhona Matshaya - Nelson Mandela University

Eskom as the main electricity producer of South Africa is facing challenges with insufficient generating capacity and capacity loss due to malfunctioning old equipment and due to increasing demand for electricity because of the population growth. This study focuses on unplanned capacity loss, which is the result of reducing the output or shutting down (or slowing down) of a generator when a reading from a SCADA sensor hits or approaches a cut-off point. These failures are highly undesirable as they result in a substantial reduction in electricity energy output. The purpose of this study is to investigate the possibility of predicting these approaching failures by modelling the data from previous failures and to look at the increasing risk of capacity loss so that the generators can be maintained and repaired before these shutdowns can occur. The use of a successful model may reduce the impact of unplanned capacity loss and make smooth operation of staff and equipment achievable.

Keywords: Eskom generation

Coauthors: I Litvine (Nelson Mandela University)

Stream: Young Statisticians

A robust approach to assessing size and value of the JSE

Zukisa Mbava - Nelson Mandela University

Empirical research has provided contradictory results when assessing whether size and value are significant predictors of expected returns. The contradictions are due to slight variations in portfolio construction, time period evaluated, and different markets investigated, amongst other things. These contradictions raise possible accusations of data-snooping. This study proposes an alternative method of assessing size and value which is considered to be robust against data-snooping.

Keywords: Returns, Size, Value, Turnover

Coauthors: S Janse van Rensburg (Nelson Mandela University) and G Sharp (Nelson Mandela University)

Stream: Financial and Business Statistics

A Bayesian Change Point Analysis of South African Financial Time Series

Siyabonga Mbonambi - Nelson Mandela University

The study endeavours to identify sudden changes in South African financial time series, and to investigate the cause of these sudden changes. In particular, the JSE/FTSE All Share Index will be investigated. This objective is set to be achieved using the Bayesian change point analysis technique. Given sudden and unexpected changes in the economic, political and social situations in the country, it is of interest to observe whether these changes affect financial time series. A study of the change point phenomena will enable us to identify the locations of these sudden changes. The occurrence of the sudden changes within financial time series could have a large impact on investors and government policy makers; as such, an understanding of the relationship between the data and the events which occur may provide domain experts with information which could assist them in decision making. In summary, the Bayesian change point technique will be used as a diagnostic tool applied in a retrospective manner.

Keywords:

Coauthors: S Das (Council for Scientific and Industrial Research and Nelson Mandela University) and S Mangisa (Nelson Mandela University)

Stream: Young Statisticians

The Kumaraswamy Log-Logistic Weibull Distribution: Model, Properties and Applications

Precious Mdlongwa - Botswana International University of Science & Technology

A new distribution called the Kumaraswamy Log-Logistic Weibull is introduced and its statistical properties are explored. This new distribution, which combines the Kumaraswamy and Log-logistic Weibull distributions, is more flexible for data modelling. Maximum likelihood is used for estimating the model parameters. Finally, an application of the model to a real data set is presented to illustrate the usefulness of the proposed distribution.

Keywords: Kumaraswamy Distribution, Log-Logistic Weibull Distribution, Maximum Likelihood Estimation

Coauthors: BO Oluyede (Georgia Southern University) and AKA Amey (Botswana International University of Science & Technology)

Stream: Young Statisticians

An application of a mixture of logistic regressions

Sollie Millard - University of Pretoria

Logistic regression is frequently used in industry for propensity based modelling. This talk gives an introduction to the use of mixtures of logistic regressions to model a specific event. The approach allows for the clustering of observations into different component logistic regression models based on the underlying latent behavioral structure. We illustrate the use of the EM algorithm to perform maximum likelihood estimation of the parameters.

Keywords: Mixture regression; logistic regression; EM algorithm

Coauthors: Frans Kanfer (University of Pretoria)

Stream: Applied Statistics and General

Sparse convex optimisation to solve a Sudoku problem

Jacob Modiba - University of Pretoria

Sparse estimation methods are aimed at using or obtaining parsimonious representations of data or models. Optimization is seeking values of a variable that lead to an optimal value of the function that is to be optimized. Suppose we have a system of equations where there are more unknowns than the equations. This type of system has infinitely many solutions. If one has prior knowledge that the solution is sparse this problem can be treated as an optimization problem. We will discuss the convex algorithms for finding the sparse solution. We use convex algorithms since they are relatively easy to implement. The classes of methods discussed are convex relaxation, greedy algorithms and iterative thresholding. We will then compare these algorithms by applying them to a Sudoku problem.

Keywords: Optimisation, sparse estimation

Coauthors: I Fabris-Rotelli (University of Pretoria)

Stream: Young Statisticians

Survival Analysis of HIV/AIDS Patients in the Limpopo Province, South Africa

Khehla Daniel Moloi - University of Limpopo

HIV/AIDS is a major concern throughout the world. South Africa has the largest anti-retroviral treatment (ART) programmes globally. Despite all these efforts HIV prevalence remains high among the general population, although it varies markedly between provinces. The Limpopo province is one of the provinces in South Africa with high HIV/AIDS prevalence. Even though ART improves the survival rate of HIV/AIDS patients, a number of factors affect the time to death of patients who are under ART. This study focuses on exploring and identifying potential risk factors for the survival rate of HIV/AIDS patients in the Limpopo province. Both semi-parametric and parametric survival models were fitted to the secondary HIV/AIDS dataset which was obtained from the Limpopo Department of Health. Significantly higher risk of death was observed for clinical stage 4 as compared to the other three clinical stages. The hazard rate for HIV/AIDS patients who received Tuberculosis treatment before being diagnosed with HIV is also significantly higher than in patients who were free of TB. The hazard of death of HIV/AIDS patients in the Mopani district is almost three-times higher than of patients who are living in the Waterberg district. Moreover, the survival rate was found to be significantly lower in specialised health care facilities than in other health centres.

Keywords: Survival, HIV/AIDS, hazard, semi-parametric

Coauthors: YG Kifle and K Zuma

Stream: Biostatistics

Applications of Statistical Learning Theory Towards Prediction of Student Success

Moeketsi Simon Mosia - University of The Free State

Student success is a topic of interest for institutions of higher learning, and the reputation of an institution of higher learning, amongst others, depends on its students' throughput rate. Furthermore, in countries like South Africa universities' funding by the government is mainly determined by the students' throughput rate. Thus the ability to predict students' success has become important for reasons which include the ability to identify at-risk students for the purpose of providing adequate student support. The purpose of this paper is to contribute towards an ongoing research project on learning analytics which seeks to predict student success. The current paper applied statistical learning theory to build a predictive model of students' success for a first year module at a South African university. Data used in this study were students' online assessments such as discussions, quizzes, and assignments and students' profile. The preliminary results of this study reveal that students' online assessments were positively correlated to student success. In the same vein the results of the study further reported a fairly built approximate model to predict student success.

Keywords: Student success, statistical learning theory, ridge regression, Lasso regression

Coauthors:

Stream: Statistics in Education

Availability analysis of systems in the presence of common cause and human error subject to variant hardware repair time models

Paul Mostert - Stellenbosch University

The availability analysis of a two-unit system of two different configurations (series and parallel) subject to unit failures, common cause failures and human error is studied when the repair time of hardware failure is constant hypo-exponential, 2-stage hypo-exponential or Erlang distributed. Small sample limitations require a Bayesian approach as adaptation with different types of priors to be assumed for the unknown parameters to evaluate the availability of the two-unit system. Gibbs sampling and the Metropolis Hastings algorithm are used to derive the posterior distribution for the steady state availability. The performance of the estimators is evaluated in terms of bias, mean squared error and frequentist coverage through a simulation study.

Keywords: Availability, Bayesian estimation, common cause failures; frequentist coverage, human error, MCMC

Coauthors:

Stream: Applied Statistics and General

Determinants of Under-five mortality in South Africa: Results from a cross-sectional Demographic Health Survey

Tshaudi Motsima - University of South Africa and Tshwane University of Technology

Under-five mortality is a public health challenge in South Africa and other developing countries. The aim of this study was to identify factors associated with under-five mortality in South Africa, taking into account clustering. SADHS data of 1998 were used. From a total of 12 247 households, 11 735 women aged 15-49 years were interviewed. Information about their reproductive health, child survival status and socio-economic status was collected. Survival analysis techniques were used to understand under-five mortality and its determinants. A frailty model incorporating family and community frailty effects was implemented. Children of multiple birth (HR: 4.15; $p = 0.000$), children who were never breastfed (HR: 3.89; $p = 0.000$), children from households without access to piped water (HR: 1.52; $p = 0.027$) and children who stayed in informal dwelling units (HR = 1.77; $p = 0.000$) were significantly at higher risk of dying before their fifth birthday. Family frailty effect was highly significant ($\theta = 1.84$; $p = 0.000$). The results revealed that type of birth, breastfeeding, access to quality water and type of dwelling unit were significant determinants of under-five mortality. The findings further confirmed that children belonging to the same family and children belonging to the same community share certain unobserved characteristics that put them at risk of death.

Keywords: Family frailty, community frailty, frailty effect, under-five mortality, determinants

Coauthors: K Zuma (Human Science Research Council)

Stream: Biostatistics

Analysis of road traffic accidents in the Limpopo province using Generalized Linear modelling

Peter Mphekgwana - University of Limpopo

Death and economic losses due to Road Traffic Accident Deaths (RTADs) are global public health and developmental problems. We investigate factors that contribute to RTADs in the Limpopo Province using prediction models such as Generalized Linear Models (GLM's) and zero-inflated models. Initial results from the analysis of RTADs are presented and analytic issues are discussed. First, we present a preliminary analysis of the road traffic accident and death data to determine its distributional properties. We then present results of fitting a standard Poisson distribution to the secondary data and compare this model with negative binomial and zero-inflated models.

Keywords: Count Models

Coauthors: A Tessera (University of Limpopo) and N Yibas (University of Limpopo)

Stream: Young Statisticians

Improvement on the imputation model methodology for missing expenditure for the South African R&D survey data

Precious Mudavanhu - Human Sciences Research Council

Survey data is commonly faced with a problem of missing data. The South African Research and Development survey team uses the cold deck imputation to mitigate the problem of missing data during data analysis. Cold deck imputation is a statistical procedure that is used to replace the value of an item with a constant value, usually from an external source such as a value from a previous survey. This study seeks to improve on the methodology currently used to impute the missing R&D expenditure data. In the current process the missing values for R&D expenditure are imputed with the previous year's R&D expenditure values adjusted by an overall annual GDP inflation factor. The study seeks to improve this methodology by adjusting the previous R&D expenditure values using sectoral GDP inflation factors for each entity that falls in that specific sector. The accuracy of the proposed methodology is assessed by cross validating the imputed R&D expenditure values against 1) the actual values supplied by the respondent and 2) the values generated by the current model. A conclusion on the performance of the proposed methodology is drawn and a recommendation made on whether or not to adopt it.

Keywords: Expenditure, imputation

Coauthors: M Sithole (Human Sciences Research Council)

Stream: Applied Statistics and General

Dealing with missing data when modelling multi-state panel data

Chris Muller - Stellenbosch University

Medical data often take the form of panel data, where the same measurements are taken from the same patients over a specified time period. In HIV/AIDS studies these measurements include covariates that can be used to classify each patient into a distinct disease state. The transition of patients through these disease states can indicate the effectiveness of a given treatment. Unfortunately, as with most data measured over time, missing observations may occur. In this talk methods of dealing with missing data in panel datasets are investigated and compared.

Keywords: Panel Data, Multi-State, Markov Model

Coauthors: PJ Mostert (Stellenbosch University)

Stream: Biostatistics

A Comparative Investigation in the Analysis of Malaria Re-Infected Patients using Accelerated Failure Time Models and Cox Proportional Hazards Models

Ruffin Mutambayi - University of Fort Hare

Malaria is an infectious disease caused by the Plasmodium parasite, and is one of the most important causes of mortality globally. This study assesses the prototypes of malaria re-infection of patients after treatment, to determine the risk factors for re-infection. The study includes 109 malaria outpatients in Lubumbashi who had re-infection status after a 180-day follow-up. Authors used the Akaike Information Criterion and Cox-Snell Residuals to analyze the re-infection data. The results from the analysis show that Gamma was better in the analysis compared to accelerated failure-time in the study. Although many researchers prefer the proportional hazard model in analysis of survival data, the accelerated failure-time model is a good alternative as it does not require proportionality of hazards as a key assumption.

Keywords: Accelerated failure time model, Cox proportional hazards model, Re-infection, Malaria

Coauthors: Ruffin Mutambayi (University of Fort Hare), Qin YongSong (University of Fort Hare) and Azeez Adeboye (University of Fort Hare)

Stream: Biostatistics

Polychotomous Logistic Regression Exploration of Determinants of Academic Achievement in the Lead-in Statistics Module: A Case Study of Undergraduate Students at the University of Fort Hare

Ruffin Mutambayi - University of Fort Hare

This paper explores the determinants of undergraduate students' academic performance at UFH. Attention was focused on attempting to corroborate the association linking gender, age, residence, faculty of study, entry points, progenitors' social-economic standing, previous school grounding, class attendance, study groups to the academic achievement of UFH's students using multiple logistic concepts. A multiple logistic regression model comprising all the predictors considered in this research is fitted and a statistical significance test of individual coefficients is conducted. On the one hand, the Wald test indicates the statistical significance of covariates residence of students, faculty of study and social economic status of students. On the other hand, the covariates gender, age, admission points, parent's social status, former school background, class attendance and study group were found not to be significant. Conclusively, a multiple logistic regression prototype that incorporates statistically significant covariates only was modelled. The researchers concluded that the full model was fitting the observations based on the comparison of the all-inclusive descriptive power and the Receiver Operator Characteristic related to the full and the reduced models.

Keywords: Multiple Logistic Regression, Analysis, Factors

Coauthors: M Mutambayi (University of Fort Hare), Q YongSong (University of Fort Hare) and A Adeboye (University of Fort Hare)

Stream: Statistics in Education

Contributions to κ - μ models in wireless systems

Priyanka Nagar - University of Pretoria

The κ - μ distribution is a popular model for the fading observed due to clusters of obstacles between a transmitter and a receiver in wireless communication channels. This paper relaxes the assumption of normality and allows for heavy tailed behaviour under the elliptical law. It is demonstrated that by adopting the t-distribution, a lower outage probability is obtained.

Keywords: Elliptical class, fading, outage probability, t-distribution

Coauthors: A Bekker (University of Pretoria) and M Arashi (Shahrood University of Technology)

Stream: Young Statisticians

Space time pattern mining of matric pass rates

Arulsivanathan Naidoo - Statistics South Africa

This paper will combine the school pass rates with the sub place data for the Census 2011 to show firstly a point pattern analysis of the performance of schools. The pass rates of the schools show a positive Moran autocorrelation indicating a clustering of school marks. A Getis - Ord hot spot analysis identifies the high performing school cluster as well as the low performing schools. The Anselin outlier analysis identifies a high performing school among a cluster of low performing schools. An OLS regression analysis was performed to identify Census attributes that contribute to school performance. The residuals for the OLS regression model are found to be autocorrelated, indicating that the global OLS model is not a good fit. A Geographically weighted regression model was fitted which showed an increase in the R-square value and a decrease in the AIC value indicating a better fit. Finally, the past 10 years' school pass rates were analysed using the space time cube and emerging hot spot analysis. The Mann Kendal Statistic was used to differentiate between categories of hot spots. With the resultant trend z-score and p-value for each location with data, and with the hot spot z-score and p-value for each bin, the Emerging hot spot analysis tool categorizes each study area location as either new hot spot, consecutive hot spot, intensifying hot spot or persistent hot spot.

Keywords: Space time cube, emerging hot spot analysis

Coauthors:

Stream: Official Statistics

Modeling the impact of climate variability on diarrhoea-associated diseases in Zambia (2010-2013)

Jason Nakaluudhe - University of Namibia

This study describes the relationship between climatic variables and diarrhoea cases in two districts

(Livingstone and Ndola) of different agro-ecological zones of Zambia. The study used climatic and aggregated monthly surveillance diarrhoea incidence data of four consecutive years (2010–2013). Both Poisson and Negative binomial regression models were used to assess the sensitivity of disease to climatic variability. Three climatic variables: minimum temperature, maximum temperature, and rainfall, at varying lags were employed. Non-linearity, to capture seasonality and long-term trends, was explored using a semi-parametric predictor. A total of 314,102 cases (Livingstone: 75,350; Ndola: 238,752) of diarrhoea occurred in two districts of Zambia. Results from Poisson regression indicated that minimum temperature (at lags 1, 2 and 3), maximum temperature (at lags 1, 2 and 3) and rain (at lags 1, 2 and 3) were strongly related to the incidence of diarrhoea in Ndola. In Livingstone, minimum temperature (at lags 1), maximum temperature (at lag 1, 2 and 3) and rain (at lag 1 and 2) were associated to diarrhoea. The association between climatic variables and diarrhea incidence has profound implications, considering worsening climate change. The risk seems to be more pronounced in arid areas of Zambia; as these become more wet, the risk increases. The results from a Negative binomial regression did not indicate any better model, probably because data was not overdispersed.

Keywords: Temperature; diarrhoea; morbidity; climate change; negative binomial regression model

Coauthors: L Kazembe (University of Namibia), F Masaninga (WHO Zambia) and P Songolo (WHO Zambia)

Stream: Biostatistics

Application of Random Survival Forests in Understanding the Determinants of Under-Five Child Mortality in Uganda

Justine Nasejje - University of KwaZulu-Natal

Uganda, like all other Sub-Saharan African countries, has a high under-five child mortality rate. To inform policy on intervention strategies, sound statistical methods are required to identify factors strongly associated with under-five child mortality rates. The Cox proportional hazards model has been a common choice in analyzing data to understand factors associated with child mortality rates, taking age as the time-to-event variable. However, due to its restrictive proportional hazards (PH) assumption, some covariates of interest which do not satisfy the assumption are often excluded from the analysis to avoid misspecifying the model. Survival trees and random survival forests are becoming popular in analyzing survival data particularly in the case of large survey data, and could be attractive alternatives to the Cox model. We adopt random survival forests which have never been used in understanding factors affecting under-five child mortality rates in Uganda using Demographic and Health Survey (DHS) data. Thus the first part of the analysis is based on the use of the classical Cox PH model and the second part of the analysis is based on the use of random survival forests.

Keywords: Cox proportional hazards model, proportional hazards assumption, Survival trees, Random survival forests

Coauthors: Henry Mwambi

Stream: Biostatistics

Prediction Error Estimation of the Survey-Weighted Least Squares Model under Complex Sampling

Ariane Neethling - Stellenbosch University

Many large-scale surveys use a complex sampling design where each observation unit is assigned a sampling weight which is developed over different stages. Survey-weighted least squares modelling (SWLS), the linear modelling of a continuous response based on its relationship with a number of covariates, correctly accounts for this complex sample design. When doing statistical modelling it is important to determine how well the selected model performs in the prediction of a future response. Cross-validation methods have long been used for this purpose under i.i.d data modelling, but not for the modelling of CS data. This talk introduces cross-validation methods for the evaluation of SWLS models. The performance of the different cross-validation methods as prediction error estimation tools is evaluated by simulation.

Keywords: Cross-validation, Sampling weights, Benchmarking, Inference
Coauthors: R Luus (University of the Western Cape) and T de Wet (Stellenbosch University)
Stream: Applied Statistics and General

Threshold autoregressive (TAR) time series models

Ané Neethling - University of Pretoria

Nonlinear time series models have become popular in recent years, especially in economics and finance. Economic theory often includes the idea that there will be a change in behaviour if some change occurs in another variable. These upswings and downswings in the economy are referred to as regimes. Standard linear time series models, such as the autoregressive moving average (ARMA) processes, are not applicable in the presence of regime-switching behaviour. We investigate the modelling of nonlinear time series processes. In particular, we focus on the threshold autoregressive (TAR) time series process with application to a South African data set.

Keywords: Nonlinear time series model; regime-switching; testing for nonlinearity
Coauthors: PJ van Staden (University of Pretoria)
Stream: Young Statisticians

A point process characterisation of extreme temperatures

Murendeni Maurel Nemukula - University of Limpopo

The point process modelling approach is considered an elegant alternative to extreme value analysis, because of its ability to model both the frequencies and intensity rates of the occurrence of extremes. We apply point process modelling (both stationary and non-stationary) to Average Maximum Daily Temperature in South Africa. A penalized regression cubic smoothing spline function is used for non-linear detrending of the data and for determination of a fixed threshold above which excesses are extracted and used. An extremal mixture model is then fitted to determine a threshold in which a boundary corrected kernel density is fitted to the bulk model and a generalized Pareto distribution fitted to the tail of the distribution. The data exhibits properties of short-range dependence and strong seasonality, leading to declustering. An interval estimator method is used for declustering data for the purpose of fitting point process models to cluster maxima. Models used in this paper are nested and likelihood ratio tests are conducted using the deviance statistic. The reparameterization approach is used for determining the frequency of the occurrence of extremely hot days, which are found to occur 15 times per year. The modelling framework and results of this paper are important to Eskom since it is during the non-winter period that they plan for the maintenance of the power plants.

Keywords: Extreme value theory, point process, temperature.
Coauthors: C Sigauke (University of Venda)
Stream: Time Series and Stochastic Processes

Competing approaches to the visualisation of incomplete categorical data sets

Johané Nienkemper-Swanepoel - Stellenbosch University

An investigation into the optimal visualisation of categorical data sets with missing values will be presented. Visualising incomplete data allows one to explore the effect of the unobserved information on the interpretable information. The visualisation of incomplete categorical data sets using subset multiple correspondence analysis (sMCA) is compared to using GPABin for combining visualisations from plausible multiple imputations of an original incomplete data set. The GPABin approach is an amalgamation of generalised orthogonal Procrustes analysis (GPA) and Rubin's rules (-bin) for multiple imputed datasets to obtain a final visual representation of multiple imputations. A simulation study considering various data scenarios has shown promising measures of fit within the Procrustes framework, by comparing the incomplete visualisations with the MCA biplot of the simulated complete data. An overview of the simulation study will be presented with focus on the presentation of an application. Non-responses can be evaluated by visual exploration of the missing category levels only. This enables the recognition of non-response patterns which could aid in the selection of suitable imputation methods by determining the missingness mechanism from sMCA biplots. Simulated missingness mechanisms are visually confirmed from sMCA biplots and can be

used to determine mechanisms in real applications.

Keywords: Biplots, incomplete categorical data, subset multiple correspondence analysis, Procrustes analysis

Coauthors: S Gardner-Lubbe (Stellenbosch University) and NJ le Roux (Stellenbosch University)

Stream: Multivariate Data Analysis

Preparing High School Learners for a Career in Statistics through Appreciation of Basic Data Analytics

Delia North - University of Kwazulu-Natal

Statistics curricula, and teaching practices of statistics at school level, have received much attention around the world in the last two decades. The digital age has led to an increased demand for statistics analytics skills in the workplace, so that it is vital that school leavers have an appreciation of the discipline of statistics and are aware of the role of data analytics in the workplace. Consequently, there is a need to ensure that basic level statistics training at school is in tune with advocacy for the discipline, i.e. that school leavers are not just statistically literate, but have an appreciation for the power of statistics in the workplace in general when they enter institutions of higher education and make career choices. This talk focuses on some recent advances in the teaching of statistics at school level, and in particular, discusses a pilot project that aims to modernize the teaching of statistics at school and at service course level.

Keywords: Statistics Capacity Building; Job ready graduates; data analytics

Coauthors: T Zewotir (University of Kwazulu-Natal)

Stream: Statistics in Education

Approaches for Handling Time-Varying Effects in Survival and Longitudinal Models

Onyekachi Esther Nwoko - University of Cape Town

Survival models are used in analysing time to event data, which is common in medical research. The Cox proportional hazard model is commonly used in analysing time to event data. However, this model is based on the proportionality hazard assumption. Violation of this assumption leads to biased results and inferences. Once non-proportionality is established, there is a need to consider time-varying effects of the covariates. Several models have been developed that relax the proportionality assumption making it possible to analyse data with time varying effects of both baseline and time-updated covariates. I present various approaches for handling time varying covariates and effects in time to event models. They include the extended Cox model which handles exogenous time-dependent covariates using the counting process formulation by Andersen and Gill. Another is the Aalen additive model, which accounts for time-varying effects. However, there are situations where not all the covariates of interest have time-varying effects. Hence, the semi-parametric additive model may be useful. The Cox-Aalen model is an alternative model which combines Cox proportional hazards model for covariates with constant effects and the Aalen additive model for time-varying effects in a single model. Other models that will be considered are joint models for the longitudinal and time-to-event processes and a dynamic path model. These models will be applied to an HIV/AIDS dataset from SA.

Keywords: Survival models, longitudinal models, Cox proportional hazard model, time-dependent covariates, time-varying covariates

Coauthors: Francesca Little (University of Cape Town).

Stream: Biostatistics

On the mixture of Generalized Poisson and Generalized Inverse Gaussian distributions

Jamiu Olumoh - American University of Nigeria

While the negative binomial (NB) and the zero-inflated negative binomial (ZNB) distributions have been used to model count data with high dispersion and excess zeros, the joint problem of excess zero and heavy/long tail, which characterizes a number of modern datasets, has been rarely addressed. We propose a model based on the mixture of the generalized Poisson and the generalized-inverse Gaussian distributions through the framework of the Lagrangian Probability distribution. We apply the model to several different datasets including the dataset of insurgency attacks in the North Eastern part of Nigeria.

Keywords: Mixture model, Lagrangian probability distribution, generalized Poisson distribution, generalized-inverse-Gaussian distribution, Count data

Coauthors: OO Ajayi (American University of Nigeria)
Stream: Theoretical Statistics

A New Class of Log-logistic Modified Weibull Distributions with Applications

Broderick Oluyede - Georgia Southern University

A new generalized distribution called the log-logistic modified Weibull (LLogMW) distribution is presented. This distribution includes many subclasses of distributions as special cases, such as the log-logistic modified Rayleigh, log-logistic modified exponential, log-logistic Weibull, log-logistic Rayleigh, log-logistic exponential, log-logistic Weibull, Rayleigh and exponential distributions. Structural properties of the distribution including the hazard function, reverse hazard function, quantile function, probability weighted moments, moments, conditional moments, mean deviations, Bonferroni and Lorenz curves, distribution of order statistics, L-moments and Renyi entropy are derived. Model parameters of this new distribution are estimated using maximum likelihood. Finally, real data examples are presented to illustrate the usefulness and applicability of the model.

Keywords: Log-logistic Distribution; Modified Weibull Distribution; Log-logistic Modified Weibull Distribution; Probability Weighted Moments; L-Moments; Maximum Likelihood Estimation.

Coauthors: F Bindele (University of South Alabama), B Makubate (Botswana International University of Science and Technology) and Shujiao Huang (BB&T Bank)

Stream: Theoretical Statistics

Goodness-of-fit test for multivariate distributions based on the Mahalanobis distance

Wallina Oosthuizen - University of the Free State

McAssey (2013) considered an empirical goodness-of-fit test for multivariate distributions, based on the Mahalanobis distances. In this presentation we expand the approach of McAssey used for continuous multivariate distributions to a discrete multivariate setting. After fitting a distribution to the Mahalanobis distances, the Shannon entropy is used to evaluate the appropriateness of the Mahalanobis measure for describing multivariate data. The test statistic used by McAssey will be compared to the well-known chi-squared goodness-of-fit test statistic. This goodness-of-fit test will be illustrated on multivariate Poisson data. A multivariate Poisson dataset will be constructed with the use of a copula function where a multivariate uniform distribution is used.

Keywords: Mahalanobis distance, multivariate distribution, multivariate goodness-of-fit test, Shannon entropy.

Coauthors: DJ de Waal (University of the Free State)

Stream: Multivariate Data Analysis

A LULU noise removal algorithm for images using principal component analysis

Christine Papavarnavas - University of Pretoria

Among image smoothing methods Principal Component Analysis (PCA) provides an efficient image denoising scheme. PCA in statistical signal processing ensures that an image's local features are effectively preserved and the noise is removed. The PCA-based denoising scheme is spatially adaptive since the image is transformed by skillfully computing a locally fitted basis. Classical PCA reduces the dimensionality of a dataset, transforming the original dataset, thus preserving only the predominantly significant principal components and hence removing noise and trivial information from the image. We investigate using the DPT with PCA for noise removal.

Keywords: LULU; LPG-PCA; image analysis; denoising

Coauthors: I Fabris-Rotelli (University of Pretoria)

Stream: Young Statisticians

Variational methods in statistical inference

Ulrich Paquet -

The expectation maximization (EM) algorithm forms the basis of many methods of statistical inference in latent variable models. These models range from basic mixture models to hidden Markov models to models

of text or dynamical systems. As an energy function, and algorithm to minimize it, it forms part of a rich family of variational methods. In this workshop, we will trace the EM algorithm through the 90s, 2000s, and 2010s, to understand how it fits into a much larger context. In particular, we will look at variational inference for latent variable models, first with EM, and then with the variational Bayes (VB) framework that built upon it. Our journey through the advances in VB in the 90s and 2000s will take us to message passing algorithms and modern stochastic optimization schemes. Stepping into the deep learning world of the 2010s and possibly beyond, we will follow the EM energy function into very flexible models that incorporate latent variables in deep neural networks. Finally, we will discuss the EM algorithm and VB framework in the context of an even richer family of approximation schemes, and investigate why they have stood the test of time.

Keywords:

Coauthors:

Stream: Workshop

Statistics, everywhere

Ulrich Paquet -

In this talk, we'll take a whirlwind tour of problems that arise in very large online systems, like recommender systems in online marketplaces, search engines, and computational advertising, where statistics and probability theory play an integral role. In each, there are real metrics or key performance indicators (KPIs) that need to be optimized: click-through rates, revenue, the retention rate of users, user engagement metrics, and many more. All of these metrics guide the design of online systems, and to do that, these online systems rely on large but intricate statistical models and a healthy understanding of causality. As examples, the talk will touch on statistical models for recommender systems and for online advertising, which model users' preferences on an individual level, are scalable, and allow for predictions to be made in real time. We'll discuss the importance of making calibrated predictions, and the dangers of sweeping it under the carpet. The causal effect on KPIs when to statistical models are introduced or changed can only be tested with some flavor of A/B testing, and we'll touch on why proper measurement, A/B testing and causality cannot be ignored. Ultimately, the success of these large-scale statistical models directly effects the success of online systems. In the engines of large online systems, we'll find statistics. Everywhere.

Keywords:

Coauthors:

Stream: Plenary

The Kumaraswamy Log-logistic Poisson Distribution

Mavis Pararai - Indiana University of Pennsylvania

A new distribution called the Kumaraswamy Log-logistic Exponential Poisson (LLOGEP) distribution is introduced. Some statistical properties of the new distribution including the expansion of the density function, quantile function, hazard and reverse hazard functions, moments, conditional moments, moment generating function, skewness and kurtosis are presented. Mean deviations, Bonferroni and Lorenz curves, Rényi entropy and distribution of the order statistics are derived. Maximum likelihood is used to estimate the model parameters. A simulation study is conducted to examine the bias and mean square error of the maximum likelihood estimators and width of the confidence interval for each parameter. Applications of the model to real data sets are presented to illustrate the usefulness and importance of the proposed distribution.

Keywords: Maximum Likelihood Estimation

Coauthors: Gayan Warahena Liyanage (Central Michigan University) and Nathan Lilla (Indiana University of Pennsylvania)

Stream: Theoretical Statistics

The impact of mis-specification of the deterministic trends on the cointegration vectors: an application to the bi-variate case on South African employment and earning

Sagaren Pillay - Statistics South Africa

This paper investigates the impact of different specifications of deterministic trends in the vector error correction model form estimated with Johansen's multivariate maximum likelihood approach. Using South

African employment and earnings data we show the impact of the mis-specification of the deterministic trends on the cointegration vectors. The study suggests that great care must be exercised in model specification. The inclusion or exclusion of the deterministic trend should be clearly justified to avoid misleading results.

Keywords: Deterministic trend, cointegration vector

Coauthors:

Stream: Time Series and Stochastic Processes

Measuring Persistence in Time-Series using Paired Comparisons Judgements

Sihle Poswayo - Nelson Mandela University

Persistence is an important dynamic property of a time series as it provides an understanding of the behaviour of the series. One way to understand persistence is to see it as opposite to jaggedness. The oldest and best-known tool to measure persistence is the so-called rescaled range (R/S) analysis popularized by Mandelbrot. Since then, many methods have been developed to measure persistence. We suggest to use subjective judgements to measure persistence in time series by comparing pairs of graphs with different Hurst exponent. The group of respondents consisted of 40 volunteers who were asked to identify which of two presented graphs is more jagged (that is, less persistent). The graphs were simulated using the time series package of Mathematica. The responses were processed using an algorithm based on the Thurstone-Mosteller model for paired comparisons. The results of the analysis show that the human eye is capable of distinguishing graphs of time series with a difference in Hurst exponent as small as 0.02.

Keywords: Persistence, Hurst Exponent, Paired Comparisons.

Coauthors: Igor Litvine (Nelson Mandela University)

Stream: Time Series and Stochastic Processes

South African high school learners' understanding of research and statistics as a tool of research

Eva Rapoo - University of South Africa

UNISA's College of Science, Engineering and Technology has for several years been hosting a Learner Research Summit (LRS), a community engagement project centred on mentoring school learners in research. The Statistics Department at UNISA has an additional teaching and learning integrated goal for the project, namely to understand better the research attitudes and skills of high school learners. This will inform module development in preparation for UNISA's planned Higher Certificate of Mathematics and Statistics, which includes a research project. The LRS activities expose most participating learners for the first time to research beyond the simple projects and investigations in the high school curriculum. For the LRS, they are expected to choose their research topic and research questions, and then need to decide how to conduct the research. The following questions arise: What is their understanding of what constitutes research, and what is the purpose of research? What is their understanding of the research process and what constitutes valid research? How well are they aware of the strengths and weaknesses of various statistical tools? How far are mentors able to correct misunderstandings? We report some initial findings, compared with the literature of related findings on research skills and attitudes of school, undergraduate and postgraduate novice researchers. We also share the implications to the development of the modules for the Higher Certificate.

Keywords: High school learners, research, community engagement, module development

Coauthors:

Stream: Statistics in Education

Spatial sampling scheme for a road network

Hayley Reynolds - University of Pretoria

Waldo Tobler was quoted as saying: "Everything is related to everything else, but near things are more related to each other". This statement suggests that geographical data points exhibit correlation which is not necessarily considered when traditional sampling is used. The aim of this work is to develop a sampling scheme which accounts for eighty percent of a population. The effectiveness of traditional and spatial sampling techniques will be compared to determine which of the two is more reliable. In order to determine an optimal driving route, graph theory will be applied to the road network, where stopping points are treated

as the nodes and roads as the edges. The subjects at the selected nodes will then be sampled according to traditional as well as spatial sampling techniques. The cost of each sampling approach will be bootstrapped so as to determine whether sampling with consideration of the spatial components of the data is superior to traditional sampling approaches.

Keywords: Spatial sampling, spatial data, environmental sampling

Coauthors: I Fabris-Rotelli (University of Pretoria), T Loots (University of Pretoria) and A Stein (University of Twente)

Stream: Young Statisticians

The burden of childhood anaemia in four sub-Saharan African countries

Danielle Roberts - University of KwaZulu-Natal

Childhood anaemia is a significant public health problem faced by many developing countries, particularly in Africa. It contributes to health problems in children by affecting their cognitive and physical development, as well as affecting their immune function which can lead to increased susceptibility to infections. The causes of anaemia are multifactorial with malaria its predominant cause in East Africa, although malnutrition, micronutrient deficiencies and intestinal parasites also play a role. This study investigates the demographic, socioeconomic and geographical determinants of childhood anaemia in Kenya, Malawi, Tanzania and Uganda. The study uses data collected from nationally represented Malaria Indicator Surveys (MIS) and Demographic and Health Surveys (DHS) conducted in those four countries between 2014 and 2016. A Bayesian geo-additive model is used for its flexibility in modelling the effects of selected socioeconomic and demographic covariates, spatial effects, and the effects of clustering via the inclusion of a random effect.

Keywords: Anaemia, geo-additive model, spatial effects

Coauthors: G Matthews (University of KwaZulu-Natal), B Sartorius (University of KwaZulu-Natal) and R Snow (KEMRI)

Stream: Biostatistics

Skew-normal distributions: Advances in theory and application

Brett Rowland - University of Pretoria

The normal distribution is popular in many statistical contexts. However, due to its symmetry and tail behavior it may not be ideal to use in many real world applications. In order to alleviate the aforementioned issues, a generalised normal distribution that exhibits flexibility in its tail behavior is proposed as candidate for application of existing skewing methodology. Methods to approximate the characteristics of this new distribution, termed the skew generalised normal (SGN) distribution, and a corresponding stochastic representation are derived. The proposed SGN distribution is used in a distribution fitting context and to approximate particular binomial distributions as an application.

Keywords: Approximating binomial distribution; Skew-normal; Skew generalised normal; Stochastic representation

Coauthors: A Bekker (University of Pretoria), M Ararshi (Sharood University of Technology) and JT Ferreira (University of Pretoria)

Stream: Young Statisticians

Three Statisticians in the Karoo (to say nothing of their wives)

Robert Schall - University of the Free State

Deep in the Kro, in ONE ravine,
THREE statisticians have been seen.

3 Statisticians / 1 Ravine –

you will agree: This is extreme!

(For here, in the remote Karoo,
one would expect no more than TWO.)

You log this ratio, draw its root,

it still stays an Unlikelihood;

this poster, therefore, documents

the rarest of Extreme Events.

Keywords: Extreme Value Theory; Karoo - Ecology; Panthera Pardus - Ecology; Revolting Rhymes; Survival; Unlikelihood

Coauthors:

Stream: Applied Statistics and General

Drie Statistici innie Kro (om van hun vrouwen nog maar te zwijgen)

Robert Schall - University of the Free State

Diep in die Kro, op ÉÉN plek,

is DRIE statistici ontdek.

Dit het gebeur, dis waargeneem,

maar die getal is hoogs ekstreem!

(Want in die Kro, ver afgeleë,

verwag jy mos nooit meer as TWEE.)

Geen statistiek verklaar dié feit

se maksimum onskynlikheid;

selfs Gumbel- of Pareto-stert

is as modelle glad niks werd.

Ons doen verslag met 'n plakkaat

want – stomgeslaan – kan ons nie praat.

Keywords: Ekstreemwaardeteorie; Karoo - Ekologie; Maksimum Onskynlikheid; Oorlewing; Panthera Pardus - Ekologie; Vreeslike Versies

Coauthors:

Stream: Applied Statistics and General

An Assessment of Methods for Calculating Neural Reliability from EEG Recordings

Pieter Schoonees - Erasmus University

The neuroscience literature suggests that the level of intersubject synchronization between the neural responses of different subjects to, for example, movies is related to population-level measures of movie success, such as box office performance. Measures of such intersubject similarity are also known as neural reliability measures. The assumption is that the more engaging a naturalistic stimulus such as a movie is, the more similar the responses are even when comparing across subjects. Several studies have shown this empirically, using a variety of methods including correlation-based distance measures and component analysis techniques similar to canonical correlation analysis. We discuss these approaches and compare the existing methods to new methods for quantifying neural reliability.

Keywords: EEG, intersubject synchronization, neural reliability

Coauthors: NJ le Roux (University of Stellenbosch)

Stream: Multivariate Data Analysis

A Review On Survey Designs Adopted By Statistics Botswana

Thapelo Sediadie - University of Botswana

The National Statistics offices in Botswana and South Africa, as well as several other statistical organizations, conduct regular nation wide household surveys. These surveys are based on complex survey designs where the first stage units are selected with the inclusion probability proportional to the measure of size, and second stage units are selected according to the systematic sampling procedure. It is well known (Chaudhuri and Arnab, 1982), that the variance of the population parameters cannot be estimated unbiasedly when such a design is used. We have proposed several methods for estimating the variance of such complex survey designs. Apart from the problem of variance estimation, standard Pearsonian chi-square tests for goodness of fit, independence and homogeneity are not valid for such complex survey designs, and standard software packages SPSS, BMDP and SAS often provide erroneous results. This paper reviewed some methods of variance estimation and hypothesis testing for complex survey designs adopted by Statistics Botswana. All such methods are limited by sample size. Further research with large sample sizes is needed to find appropriate methods of variance estimation and hypothesis testing for such complex survey designs.

Keywords: A Review On Survey Designs Adopted By Statistics Botswana
Coauthors: R Arnab (University of Botswana)
Stream: Applied Statistics and General

A comparison of the EM algorithm and the method of maximum likelihood under constraints

Prenil Sewmohan - University of Pretoria

We examine and contrast the EM algorithm and the method of maximum likelihood under constraints, MLEC. We focus on maximum likelihood estimation where the density functions come from the exponential family. The two methods are compared and contrasted according to their performance and practicality in a number of applications. The applications considered here are maximum likelihood estimation in the presence of missing data in incomplete contingency tables and variance component models. We show by way of these examples and the theoretical aspects of the methods that the method of maximum likelihood under constraints is a simpler and more natural method to use than the EM algorithm particularly, in cases where there is no real missing data involved.

Keywords: Maximum likelihood, EM algorithm, MLEC, Missing data
Coauthors: HF Strydom (University of Pretoria)
Stream: Applied Statistics and General

Bayesian Methods to Impute Missing Data from Climate Variables

Siphamandla Sibiya - University of KwaZulu-Natal

Missing data are a major problem in meteorological data. The availability of a long and complete temperature record is important for carrying out a successful climatological study. However, the data in the records may contain missing values due to various reasons. To exclude cases with missing data and analyse the complete cases can result in biased estimates, reduced power and wrong inference. The objective of the study is to analyse the different methods for filling gaps in temperature data records. A Bayesian multiple imputation method is introduced to handle the missing data. In the context of the Bayesian framework, this article compares imputation under the restricted multivariate normal imputation model with the Bayesian multiple imputation method highlighting the bias and efficiency of regression standard errors in the study. We explore Multiple Imputation using multivariate normal imputation (MVNI) and the Bayesian approach through an analysis of a sample data set. Our analysis confirms that the power of Multiple Imputations lies in achieving smaller standard errors. In our sample data set the standard error was smallest for Multiple Imputation combined with Bayesian Regression.

Keywords: Missing data; Bayesian imputation; Multiple imputation; Multivariate normal imputation
Coauthors: S Ramroop, SF Melesse and P Mokilane
Stream: Time Series and Stochastic Processes

Forecasting temporal hierarchical time series: An application to South African electricity data

Caston Sigauke - University of Venda

The paper discusses an application of forecasting hierarchical electricity generated and available for distribution to the nine provinces of South Africa for the period 2002 to 2017. It is important that the forecasts are accurate and robust and that there are non-overlapping temporal aggregates at each level up to the annual level (Athanasopoulos et al., 2017). This type of forecasting is important for decision-making at different managerial levels from short-term operational to long-term strategic planning.

Keywords: Electricity demand, forecast combination, hierarchical forecasting, temporal aggregation
Coauthors:
Stream: Applied Statistics and General

Spatial effects on modelling of individual HIV status in Malawi: A generalized additive model approach

Jupiter Simbeye - University of Malawi

We investigate district unobserved spatial effects on individual HIV status in Malawi. We accomplish this through a generalized additive model applied to the 2010 Malawi Demographic and Health Survey (MDHS). The 2010 MDHS is the first of its kind to collect nationally representative HIV & AIDS data at district level for

all districts in Malawi. It contains a sample of 14, 407 adults with a successful blood test for HIV, of which 7, 090 were women (15-49 years old) and 6, 837 were men (15-54 years old). The spatial predictor was represented by the individual's district of residence, which was geo-referenced. We controlled for other demographic, behavioral and socio-economic predictors in the model. The results have shown a strong positive spatial effect, mainly among districts in the southern region of Malawi, which may have a strong cultural bearing. The study has demonstrated that when modeling geo-referenced data, it is important to account for spatial effects in order not to overestimate standard errors.

Keywords: HIV status, Generalized Additive Models, spatial effects, Malawi

Coauthors:

Stream: Applied Statistics and General

Extreme Distributions based on Process Characteristics

Ansie Smit - University of Pretoria

Extreme Value Theory (EVT) and Generalized Extreme Value Theory (GEVT) are often applied to model natural hazards such as earthquakes, floods and tsunamis by focusing only on extreme events. Weaker events, often ignored, can still provide valuable information on the underlying physical process. Most event databases contain three types of data: prehistoric, historic and the most recent, instrumentally recorded events. Prehistoric and historic datasets primarily contain only the large events but instrumental data also represent smaller events. We provide an alternative to the EVT and GEVT models for natural hazards. Likelihood functions describing the recurrence parameters of the underlying processes are defined for each of the three data types. They are combined into a single likelihood function for the whole database. Potential incompleteness, uncertainty in size of the events, and uncertainty associated with the applied recurrence model are considered. The likelihood functions for the prehistoric and historic data are modelled using extreme value distributions that are not only characteristic of the underlying process, but also linked to the parent distributions of the process. The advantage of such modelling is that the process-characteristic extreme distribution preserves parameters of the parent distributions, and as a rule, provides more accurate process information than EVT or GEVT modelling. As an example, the methodology will be applied to earthquakes in Central Italy.

Keywords: Incompleteness, uncertainty, derived extreme distributions

Coauthors: A Kijko (University of Pretoria) and A Stein (University of Twente and University of Pretoria)

Stream: Young Statisticians

Spatial Correlation of Diabetes across African Countries using Meta-Data

Adenike Soogun - University of KwaZulu-Natal

Diabetes is rising on the African continent, with high disparities between countries. However, the complexities and heterogeneity of disease levels across the continent have necessitated the need for the study. A holistic understanding of the global knowledge is important. But an accurate understanding of the disease with specific situations is essential in formulating policy and management of the disease. It is crucial to identify high-risk geographical countries and determine similar countries as regard exposure to diabetes. This study examines diabetes prevalence across the African continent with underlying socioeconomic and health risk indicators to (1) show disease mapping of diabetes prevalence and the underlying indicators and (2) to determine the geographic pattern of diabetes prevalence by identifying spatial clusters. Meta-Data was extracted from the world data bank and World Health Organization (WHO). Spatial analysis using Moran index as well as cluster analysis are implemented using R software with mapproj, mclust, ape, rgdal, spdep packages are used. Global and Local Moran's Indices were calculated to determine spatial autocorrelation in the pattern of diabetes prevalence across Africa and also according to different regions. Cluster analysis using different techniques was performed.

Keywords: Diabetes, Africa, Spatial Autocorrelation, Cluster analysis

Coauthors: S Lougue (University of KwaZulu-Natal)

Stream: Applied Statistics and General

Bayesian Hierarchical Modelling with application in Spatial Epidemiology

Richard Southey - Rhodes University

Disease mapping and spatial statistics have increased in popularity as the methods and techniques have evolved. The hyperprior sensitivity of the precision for the uncorrelated heterogeneity and correlated heterogeneity components in a convolution model will be assessed. The correlated heterogeneity will be modelled by a conditional autoregressive prior distribution and the uncorrelated heterogeneity will be modelled by a zero mean Gaussian prior distribution. The results show that the hyperprior of the precision for the uncorrelated heterogeneity and correlated heterogeneity components are sensitive to changes and will result in different results depending on the specification of the hyperprior distribution of the precision for the two components in the model. The second part of this paper will examine whether there is a difference between the proper conditional autoregressive prior and intrinsic conditional autoregressive prior for the correlated heterogeneity component in a convolution model. The results show that there is no significant difference between the results of the model with a proper conditional autoregressive prior and intrinsic conditional autoregressive prior for the South African data, although there are a few disadvantages to using a proper conditional autoregressive prior for the correlated heterogeneity. All models will be fitted in WinBUGS.

Keywords: Bayesian Statistics, Conditional Autoregressive Model, Disease Mapping, Standardised Mortality Ratio

Coauthors: L Raubenheimer (North-West University and Rhodes University) and SE Radloff (Rhodes University)

Stream: Applied Statistics and General

Spatial statistical aspects in geo-health studies for neglected tropical diseases

Alfred Stein - University of Twente

There is an increasing attention to the use of spatial and spatio-temporal studies in health studies. Such geo-health studies are in particular of interest for neglected tropical disease. In this presentation, attention will be given to aspects of cholera and diarrhea. Knowledge of the temporal trends and spatial patterns which reflect the respective roles of primary and secondary transmissions will have significant implications for effective preparedness in future epidemics. We developed generalized nonparametric and segmented regression models to describe the epidemic curve. The study showed that the epidemic rose suddenly to a peak whereas the decay was much slower. Spatial interaction occurred within 1 km radius with a maximum within 0.3 km. Significant clustering during the first week suggests secondary transmissions sparked the outbreak. The nonparametric and segmented regression models, together with the pair correlation function, contribute to understanding the respective roles of primary and secondary transmissions. In contrast, knowledge of the biological and anthropogenic characteristics of diarrhea is abundant, but little is known about their spatial pattern. In this study we observed substantial variations in the spatial distribution of the relative risk. Space-time clustering occurred in sparsely populated districts, in particular in peri-urban areas. These findings are useful public health information to complement the design of location specific interventions.

Keywords: Spatial Statistics, Health, Neglected tropical diseases

Coauthors: F Osei (University of Twente)

Stream: Biostatistics

Extreme Value-based Novelty Detection

Luca Steyn - Stellenbosch University

Novelty detection is the process of detecting when an observed event differs from the expected, normal behaviour. One approach to perform novelty detection is probabilistic novelty detection. This talk explores the use of extreme value theory to perform probabilistic novelty detection. A first extreme value-based model, termed the Winner-Takes-All, is investigated. The model's strong theoretical underpinning as well as its disadvantages are discussed. The second method reformulates extreme value theory in terms of extreme probability density. This definition is used to derive a closed-form expression of the probability distribution in the case of an underlying multivariate Gaussian probability density. It is shown that this distribution is in the minimum domain of attraction of the extremal Weibull distribution. A final method to perform novelty

detection approximates the distribution of the extreme probability density values under the assumption of a multivariate Gaussian mixture model representing the underlying probability density function. This allows novelty detection to be performed in multimodal and multidimensional settings using univariate extreme value theory. To demonstrate the application of the discussed methods a banknote authentication dataset is analysed. It is shown that extreme value-based novelty detection methods are extremely efficient in detecting forged banknotes.

Keywords: Multivariate novelty detection, extreme value theory, one-class classification

Coauthors:

Stream: Multivariate Data Analysis

Assessment of a composite index for analysing the goodness-of-fit in SEM

Carmen Stindt - Nelson Mandela University

Structural equation modelling (SEM), a statistical technique used extensively in quantitative marketing research and other domains, is an analytical approach used to model latent (unobservable) variables. Unlike distribution fitting where simple chi-squared goodness-of-fit assessment yields satisfactory results, model fit in SEM is more difficult. Researchers use several goodness-of-fit indices when motivating their model selection, and in many cases the indices used differ considerably. This study defines a composite index, combining frequently used indicators in an attempt to obtain a single index method for assessing model fit in SEM.

Keywords: Goodness-of-fit indices, model fit, structural equation modelling

Coauthors: GD Sharp (Nelson Mandela University) and M Mey (Nelson Mandela University)

Stream: Young Statisticians

Obtaining yield probabilities by using different CV%, for row/col and randomized block designs for different crops in cultivar selection

Nicolene Thiebaut - Agricultural Research Council

The selection of cultivars in the grain crop industry under different environmental circumstances is important for seed companies, farmers and industries in optimizing the profit and quality of the product. A few cultivar selection trials for different crops (maize, soy-, dry beans, wheat and sunflower) are done annually at different localities, involving high costs. Yield probabilities as a percentage of the mean yield are used as guidelines for cultivar selection in different regions and circumstances. Localities are selected for calculating yield probabilities, using a specified CV%. The CV% for the datasets are calculated for the specified row/col and randomized block designs (RBD). The presentation consists of a discussion on refining the program to obtain the results.

Keywords: Yield probability, CV%, randomized block design, row/col design

Coauthors: Annelie De Beer (Agricultural Research Council) and Deidre Fourie (Agricultural Research Council)

Stream: Biostatistics

Road extraction in remote sensing images of South African informal settlements

Renate Thiede - University of Pretoria

Informal unpaved roads are a common problem in South African municipalities. These roads are usually formed ad hoc by citizens without the knowledge of government authorities. Information on these roads is critical for sustainable city growth and maintenance. In order to estimate the number and location of informal roads within an area, the number and location of these roads should first be estimable. This paper will address a first step towards this process, namely the detection of informal roads. A method for urban road extraction proposed by Li et al. (2016) is tested on informal roads. The method relies on the hierarchical representation of the study area in a Binary Partition Tree. Regions in the image are modeled using two geometrical features, namely region elongation and compactness, and two structural features, respectively using orientation histograms and path-based morphological profiles. Path-based morphology is less restrictive than classical morphology and allows for the detection of curvilinear roads. The method is applied to two very high resolution images of areas in Northwest Province, South Africa.

Keywords: Remote sensing, road detection, binary partition tree, mathematical morphology

Coauthors: I Fabris-Rotelli (University of Pretoria), A Stein (University of Pretoria and University of Twente) and P Debba (CSIR Built Environment and University of the Witwatersrand)
Stream: Young Statisticians

Road extraction in remote sensing images of South African informal settlements

Renate Thiede - University of Pretoria

Informal unpaved roads are a common problem in South African municipalities. These roads are usually formed ad hoc by citizens without the knowledge of government authorities. Information on these roads is critical for sustainable city growth and maintenance. In order to estimate the number and location of informal roads within an area, the number and location of these roads should first be estimable. This paper will address a first step towards this process, namely the detection of informal roads. A method for urban road extraction proposed by Li et al. (2016) is tested on informal roads. The method relies on the hierarchical representation of the study area in a Binary Partition Tree. Regions in the image are modeled using two geometrical features, namely region elongation and compactness, and two structural features, respectively using orientation histograms and path-based morphological profiles. Path-based morphology is less restrictive than classical morphology and allows for the detection of curvilinear roads. The method is applied to two very high resolution images of areas in Northwest Province, South Africa.

Keywords: Remote sensing, road detection, binary partition tree, mathematical morphology

Coauthors: I Fabris-Rotelli (University of Pretoria), A Stein (University of Pretoria and University of Twente) and P Debba (CSIR Built Environment and University of the Witwatersrand)

Stream: Young Statisticians

Trees to networks: an evaluation of neural random forests

Motlamedi Thupae - University of Pretoria

This research seeks to investigate whether it is possible to restructure a collection of random forests as a collection of multilayered neural networks subject, to particular connection weights, using the R statistical software. To this end, a random forest is reformulated into a neural network, leading to new hybrid procedures, namely neural random forests. Prior knowledge of the underlying design of regression trees is used as they have less parameters than standard networks that need adjusting, as well as exhibiting less restrictions on the decision boundaries. Neural random forests consider the implications and advantages of both models and seeks to combine them in order to achieve a better performing model overall. The neural random forest uses the output of a random forest as the input for the neural network to essentially simulate a random forest (and its associated advantages) within a neural network. Neural random forests are reviewed by evaluating consistency results, numerical evidence, as well as assessments based on real data sets to gauge the method's performance against various prediction problems.

Keywords:

Coauthors: Theodor Loots

Stream: Young Statisticians

Trees to networks: an evaluation of neural random forests

Motlamedi Thupae - University of Pretoria

This research seeks to investigate whether it is possible to restructure a collection of random forests as a collection of multilayered neural networks subject, to particular connection weights, using the R statistical software. To this end, a random forest is reformulated into a neural network, leading to new hybrid procedures, namely neural random forests. Prior knowledge of the underlying design of regression trees is used as they have less parameters than standard networks that need adjusting, as well as exhibiting less restrictions on the decision boundaries. Neural random forests consider the implications and advantages of both models and seeks to combine them in order to achieve a better performing model overall. The neural random forest uses the output of a random forest as the input for the neural network to essentially simulate a random forest (and its associated advantages) within a neural network. Neural random forests are reviewed by evaluating consistency results, numerical evidence, as well as assessments based on real data sets to gauge the method's performance against various prediction problems.

Keywords:

Coauthors: Theodor Loots

Stream: Young Statisticians

A practical maturity assessment method for model risk management in banks

Liesl van Biljon - Standard Bank

We motivate and propose a qualitative method to assess the maturity of model risk management practices in banks. This method is aligned with relevant regulatory guidance and observed leading practice. It provides banks with a practical way to determine their current maturity levels with respect to model risk management practices, and to define a targeted level of maturity. It also makes clear which aspects need to be remedied to progress from the current state to the targeted maturity state. Therefore, the results of the application of this proposal provide a view of the current state of a bank's model risk management practices as well as what needs to be improved to further mitigate model risk at a targeted maturity state.

Keywords: Model risk, model risk management, financial risk models, model validation

Coauthors: Leendert Haasbroek (Standard Bank)

Stream: Financial and Business Statistics

A method for Bayesian regression modelling of composition data

Sean van der Merwe - University of the Free State

Many scientific and industrial processes produce data that is best analysed as vectors of relative values, often called compositions or proportions. The Dirichlet distribution is a natural distribution to use for composition or proportion data. It has the advantage of a low number of parameters, making it the parsimonious choice in many cases. We consider the case where the outcome of a process is Dirichlet, dependent on one or more explanatory variables in a regression setting. We explore some existing approaches to this problem, and then introduce a new simulation approach to fitting such models, based on the Bayesian framework. We cover the advantages of the new approach through simulated examples. These advantages include the ability to introduce random effects or hierarchical structures without additional complexity in the analysis.

Keywords: Compositional data, proportions, Dirichlet distribution, regression, Bayes, simulation

Coauthors:

Stream: Multivariate Data Analysis

Bayesian testing for process capability indices

Abrie van der Merwe - University of the Free State

Process capability indices have been widely used in the manufacturing industry. They measure the ability of a manufacturing process to produce items that meet certain specifications. In this talk a Bayesian method is presented for the analysis of Cpl, Cpu and Cpk. For illustration purposes an example from Chou (1994) is used. Four suppliers (processes) produce piston rings for automobile engines. The edge width of a piston ring is an important quality characteristic in the manufacturing process. A Bayesian simulation method proposed by Ganesh (2009), which is a version of Tukey's simultaneous confidence interval procedure, is implemented for the capability indices to determine which processes are significantly different from one another.

Keywords: Capability indices, Posterior distributions, Bayesian testing

Coauthors: PCN Groenewald (University of the Free State) and R van Zyl (QuintilesIMS)

Stream: Applied Statistics and General

The effect of unnecessary blocking on power in experimental design – a simulation study

Stephan van der Westhuizen - Stellenbosch University

Inclusion of a blocking factor in an experimental design is a method for incorporating underlying trends in order to reduce residual variance in an experiment. Therefore, when blocking is appropriately used in a design it may increase the power of the analysis. However, introducing unnecessary blocking factors could lead to even lower power. A simulation study is performed to see to which extent power is negatively influenced by including unnecessary blocking factors. The simulations will be performed on various power

parameters i.e. number of replicates, effect size and variance, and for various treatment designs.

Keywords: Blocking, experimental design, power analysis, simulation

Coauthors:

Stream: Applied Statistics and General

Smoothing Parameter Selection for Distribution Function Estimation using the Bootstrap

Francois van Graan - North-West University

Smoothing parameter selection of a kernel distribution function estimator is proposed using the bootstrap. The relevant criterion to be estimated is the bootstrap expected value of the mean average squared error. It is shown that this criterion is asymptotically related to mean integrated squared error.

Keywords: Distribution Function, Bandwidth, Bootstrap, Non-parametric

Coauthors:

Stream: Theoretical Statistics

The Role of the Turning Angle when Modelling Animal Movement Data using Hidden Markov Models

Bracken van Niekerk - Nelson Mandela University

Hidden Markov Models (HMMs) are good candidates for animal movement studies, as they provide a flexible modelling approach to segment the movement path into latent behavioural states, which are optimised from a sequence of successive observations. These latent states are inferred to as proxies for the behaviours of the animals. The most common metrics are the successive displacements between locations (step lengths) and turning angles (measuring the tortuosity). There has been little mention of the reasoning behind the inclusion or exclusion of the turning angle in the literature. This study investigates the statistical and ecological influence of the turning angle in terrestrial animal movement modelling when fitting HMMs. The predicted state sequence was compared for models fitted with and without the turning angle, and was measured as a percentage same state allocation. Results were obtained for GPS tracking data observed at different time scales, for both carnivore and herbivore terrestrial species, in different areas in and around Southern Africa. The inclusion of the turning angle in the modelling process is thought to overcomplicate the models unnecessarily, and in most cases does not alter the model outputs. This was predominantly found with the simpler models. The ecological interpretations of the models when including the turning angle remain mostly unchanged, or become more challenging and even irrelevant in the more complex models.

Keywords: Hidden Markov Models, turning angle, animal movement modelling

Coauthors: VL Goodall (Nelson Mandela University)

Stream: Applied Statistics and General

Feature detection using direct sampling and the Discrete Pulse Transform

Carel van Niekerk - University of Pretoria

In computer or robot vision an important concept is feature detection. Feature detection is the process of obtaining disjoint and significant features. We combine the Discrete Pulse Transform and direct sampling for this purpose. The Discrete Pulse Transform (DPT) is a digital signal processing algorithm which segments digital signals into pulses, i.e. connected components of different sizes. These pulses can be used to identify important regions in an image, and the significance of a pixel is determined by its saliency, the number of pulses which contains that feature. Direct sampling is a stochastic simulation algorithm which simulates by sampling from a training image. The process it uses is to scan through the training image and find a pixel whose neighborhood matches that of the pixel it wants to simulate best. The ability to add conditioning data to the grid of pixels to simulate potentially gives this algorithm the ability to simulate a object, such as a ball, from the training image it is given.

Keywords: LULU smoothers, DPT, image processing

Coauthors: I Fabris-Rotelli (University of Pretoria) and J van Niekerk (University of Pretoria)

Stream: Young Statisticians

The Assessment of Multiple Imputation

Michael von Maltitz - University of the Free State

The validity and efficiency of multiple imputation (MI) is defined and assessed in terms of estimate bias reduction and standard error correction when moving from analyses on incomplete data to analyses on MI-completed data. As a result, when testing new MI methods one is often asked to compare results from various analyses made on both the incomplete and completed data sets (means, percentiles, and regression models, for example). However, this is counter-intuitive, as MI was developed to dissociate the analysis task from the imputation task; imputers should be able to publish completed data for public use without knowing what the data will be used for. This paper explores the possibility of developing an assessment protocol for new MI methods that does not depend on any post-imputation modelling.

Keywords: Multiple Imputation, Sequential Regression Multiple Imputation, Incomplete Data, Simulation

Coauthors:

Stream: Multivariate Data Analysis

On the Kumaraswamy-generalised normal distribution

Matthias Wagener - University of Pretoria

We introduce the Kumaraswamy-generalised normal distribution. The distribution function of the proposed distribution is obtained by evaluating the $F(G)$ where F and G respectively denote the distribution functions of the Kumaraswamy and generalised normal distributions. Some statistical properties of the newly derived distribution are studied and the distribution is fitted to observed levels of a certain protein in athletes.

Keywords: Generated distributions, Kumaraswamy distribution, generalised normal distribution

Coauthors: S Makgai (University of Pretoria), IJH Visagie (University of Pretoria) and Andriette Bekker (University of Pretoria)

Stream: Young Statisticians

On the Kumaraswamy-generalised normal distribution

Matthias Wagener - University of Pretoria

We introduce the Kumaraswamy-generalised normal distribution. The distribution function of the proposed distribution is obtained by evaluating the $F(G)$ where F and G respectively denote the distribution functions of the Kumaraswamy and generalised normal distributions. Some statistical properties of the newly derived distribution are studied and the distribution is fitted to observed levels of a certain protein in athletes.

Keywords: Generated distributions, Kumaraswamy distribution, generalised normal distribution

Coauthors: S Makgai (University of Pretoria), IJH Visagie (University of Pretoria) and Andriette Bekker (University of Pretoria)

Stream: Young Statisticians

Sufficient dimension reduction in regressions

Lixing Zhu - Chair Professor of Statistics, Department of Mathematics, Hong Kong Baptist University

In this workshop, we will give the basic ideas and challenges of regression modelling in high-dimensional paradigms and show the necessity of dimension reduction. Afterwards, we present four popularly used methodologies in the literature. Thus, the workshop consists of three lectures: 1. Introduction to Dimension reduction; 2. Dimension reduction estimation: can linear methods solve nonlinear problems? 3. Estimation: can inverse regression methods solve forward regression problems?

Keywords:

Coauthors:

Stream: Workshop

Order determination for large dimensional matrices

Lixing Zhu - Chair Professor of Statistics, Department of Mathematics, Hong Kong Baptist University

This talk describes two longstanding problems in the model dimensionality (order) when criteria that are based on eigen-decomposition of target matrices are used in practice. First, due to the existence of some dominating eigenvalues compared to other nonzero eigenvalues, the true dimensionality is often underestimated. Second, the estimation accuracy of any existing method often relies on the uniqueness of minimum/maximum of the criterion. Yet, it is often not the case particularly for the models that converge to

a limit with smaller dimensionality. To alleviate these difficulties, we propose a thresholding double ridge ratio criterion. Unlike all the existing eigendecompositionbased criteria, this criterion can define a consistent estimate even when there are several local minima. This generic strategy is readily applied to many fields. As the examples, we give the details about dimension reduction in regressions with fixed and divergent dimensions; about when the number of projected covariates can be consistently estimated, when cannot if a sequence of regression models converges to a limiting model with fewer projected covariates; about ultra-high dimensional factor models and about spiked population models. Numerical studies are conducted to examine the finite sample performance of the method.

Keywords:

Coauthors:

Stream: Plenary