SASA

South African Statistical Association

UNIVERSITEIT
iYUNIVESITHI
STELLENBOSCH
UNIVERSITY

100
1918 · 2018

1 – 3 December 2021

SASA2021

62nd Annual Conference of the South African Statistical Association

VENUE | Stellenbosch University & Online

Abstracts

Thank you to our sponsors

SASA

South African Statistical Association

UNIVERSITEIT
iYUNIVESITHI
STELLENBOSCH
UNIVERSITY

Department of Statistics and Actuarial Science
School for Data Science and Computational Thinking

EXPLORE

DATA SCIENCE
ACADEMY

Keynotes

Stress testing behavioural and macroeconomic risks in credit portfolios

Jonathan Crook and Viani Djeundje

University of Edinburgh Business School, UK

Corresponding author: j.crook@ed.ac.uk

Large banks are required to stress test their credit portfolios annually under Basel II. Stress testing credit portfolios to macroeconomic shocks at account level involves parameterising a model predicting probability of default or default rates followed by hypothesising specific shocks, or by simulation to derive a value at risk (VaR) or expected shortfall (ES) 12 months into the future. But the probability of default is also correlated with time varying behavioural variables, which in turn are correlated with the macroeconomy. Simulation studies have estimated the VaR when mutually consistent macroeconomic values have been simulated or when behavioural variables have been simulated but not when both are simulated. In this paper we present a method to simulate both behavioural and macroeconomic variables 12 months into the future whilst maintaining the correlation structure between them to derive a more comprehensive simulation methodology to stress test a credit portfolio.

Ethical Machine Learning in Managing a Health Pandemic

McElory Hoffmann and Johann van der Merwe

Praelexis, Stellenbosch, South Africa

Corresponding author: johan@praelexis.com

Efficient acquisition and processing of information is a key factor in controlling epidemics. Relevant elements of the health status (e.g. infection state) of individuals or groups of individuals, mobility patterns of citizens, or information on social contacts and social networks are examples of such information. Information processing for decision making may include insights from machine learning models for the analysis of the dynamics of an epidemic using and linking such data. However, privacy interests and other ethical issues have to be carefully considered. In this presentation Drs Hoffmann and Van der Merwe will give an overview of the application of machine learning in an interdisciplinary study where Praelexis is the industry partner, as well as the "Ethical User Story" approach they developed to address ethical concerns.

Irrational Exuberance: Correcting Bias in Probability Estimates

Gareth James

University of Southern California, USA

Corresponding author: gareth@marshall.usc.edu

We consider the common setting where one observes probability estimates for a large number of events, such as default risks for numerous bonds. Unfortunately, even with unbiased estimates, selecting events corresponding to the most extreme probabilities can result in systematically underestimating the true level of uncertainty. We develop an empirical Bayes approach “Excess Certainty Adjusted Probabilities” (ECAP), using a variant of Tweedie’s formula, which updates probability estimates to correct for selection bias. ECAP is a flexible non-parametric method, which directly estimates the score function associated with the probability estimates, so it does not need to make any restrictive assumptions about the prior on the true probabilities. We demonstrate that ECAP can provide significant improvements over the original probability estimates.

GEMS: Supporting Data-Driven Agri-Food Innovation from Molecules to Markets

Ali Joglekar

University of Minnesota, USA

Corresponding author: joglekar@umn.edu

Unlocking the promises of Big Data requires a clear vision, a talented and agile cross-disciplinary team, and an agile but long-term commitment to developing the tools and systems required to tackle the real world challenges of turning data into actionable information. This is the challenge G.E.M.S (which stands for Genetics x Environment x Management x Socioeconomics) — an international agroinformatics initiative jointly led by the College of Food, Agricultural and Natural Resources Sciences (CFANS) and the Minnesota Supercomputing Institute (MSI) at the University of Minnesota — is tackling head on. GEMS is the first and the only system designed from the very start to support and functionally integrate spatially and temporally distributed genomic, environmental, management, and socioeconomic data in a single integrated platform. Making diverse data interoperable is only the start. With a focus on effective data security, flexible data sharing, and environments suitable for advanced and ever-improving data analytics, GEMS is addressing the hard problems that until now have kept most from unlocking the power and potential of the Big Data revolution.

Incorporation of historical information in the analysis of current data: A review of Bayesian methods with applications in pharmaceutical research

Emmanuel Lesaffre

KU Leuven, Belgium

Corresponding author: emmanuel.lesaffre@kuleuven.be

One of the key features of the Bayesian approach is the possibility to formally include prior information into the analysis of current data. This prior information can come from expert knowledge or from historical data or a combination of both. In this talk we focus on how to include prior information from historical data in pharmaceutical research. A naïve approach is to fully incorporate the historical information into the current analysis. It has been recognized that this most of often not a good idea, because it basically comes down to pooling the historical and current data. Additionally, time trends and conditions that differ between historical and current data may prevent to (fully) include the historical information into the current analysis. In this respect, Pocock (1976) has specified some (stringent) criteria in order to take historical information into account. In the last two decades formal procedures have been suggested for the inclusion of historical information, i.e. the power prior, the meta-analytic and the commensurate prior. We review in this talk especially the first two approaches for a single historical study and multiple historical studies. We will also focus on the use of historical controls. Without doubt the use of historical information is becoming increasingly important in pharmaceutical research because it allows to reduce the current sample size. But these methods are also becoming indispensable in e.g. paediatric studies where information from adult studies can be useful to obtain a clearer view on the efficacy and safety of experimental treatments in children and adolescents. Other applications where the use of historical information is important are bridging studies, trials that incorporate real world evidence and trials on medical devices. We discuss the pros and cons of the formal procedures, and we end the talk with the exploration of their frequentist properties. Applications in pharmaceutical research will illustrate the concepts and results.

The drastic Under-representation of African Researchers in Africa-related Research

Saralees Nadarajah

School of Mathematics, University of Manchester, UK

Corresponding author: Saralees.Nadarajah@manchester.ac.uk

In an ever more connected world one would expect to see collaborative efforts in academia build bridges between nations, continents and peoples. While the internet and digitisation have broken down boundaries and significantly lowered the financial obstacles as well as delays in time that came with international partnerships, collaborations seem to have been strengthened between industrialised nations. In this talk, I will analyse publication data of articles, notes and presentations on that reference the continent of Africa, as well as nations past or present on the continent and investigate the distribution of author affiliation within and outside the continent.

Contributed talks

Spatial variation in the basic reproduction number of COVID-19: A systematic review

Nada Abelatif[1], Renate Thiede[2], Inger Fabris-rotelli[1], Raeesa Manjoo-Docrat[3], Jenny Holloway[4], Charl Janse vsn Rensburg[1], Pravesh Debba[4], Nontembeko Dudeni-Tlhone[4], Zaid Kimmie[5], Alize le Roux[6] and Sibusiswe Makhanya[7]

[1]*Medical Research Council*

[2]*University of Pretoria*

[3]*University of the Witwatersrand*

[4]*Council for Scientific and Industrial Research*

[5]*Foundation for Human Rights*

[6] *Institute for Security Studies Africa*

[7] *IBM Research Africa*

Corresponding author: inger.fabris-rotelli@up.ac.za

The sudden emergence of the COVID-19 pandemic in early 2020 gave rise to an explosion of epidemiological modelling. The basic reproduction number (R_0) quantifies the average number of people that one infectious person will infect in a fully susceptible population. This was an important parameter at the start of the pandemic, as it was used to predict the course of the novel disease and guide decision-making. This systematic review aims to determine how early estimates of R_0 for COVID-19 varied across countries in the initial months of the pandemic (January - June 2020) and which estimates were available for Africa. We found that estimates of R_0 differed considerably between countries, ranging between 0.48 and 7.2 without outliers. Although developing countries mostly had fewer available estimates, these were generally lower and less variable. Estimates for Africa were sparse and produced mainly by researchers outside the continent. The spatial variability in estimates of R_0 at the start of the pandemic indicates that there was no one-size-fits-all model for the initial spread of COVID-19: this demonstrates the need for modelling at a local level. The sparsity of estimates for developing and particularly African countries and the lack of peer-reviewed papers providing early R_0 estimates from African-based researchers is concerning, as research during the early stages of an outbreak is critical for mitigating disease spread.

Forecasting Volatility in Commodity Markets with Long-Memory Models

Mesias Alfeus[1] and Christina Nikitopoulos[2]

[1] *Stellenbosch University, Stellenbosch, South Africa*

[2] *University of Technology Sydney (UTS), Sydney*

Corresponding author: mesias@sun.ac.za

Commodities are the most volatile markets, and forecasting their volatility is an issue of paramount importance. We examine the dynamics of commodity markets volatility by employing three typical long-memory models: fractional integrated generalized autoregressive conditional heteroscedastic (FIGARCH), fractional stochastic volatility (FSV) and heterogeneous autoregressive (HAR) models. Based on a high-frequency futures price dataset of 22 commodities, we confirm that the volatility of commodity markets is rough, and volatility components over different horizons are economically and statistically significant. Long-memory with anti-persistence is evident across all commodities, with weekly volatility dominating in most commodity markets and daily volatility for oil and gold markets. HAR models display a clear advantage in forecasting performance compared to the two other models for short horizons, while fractional volatility models yield comparative better forecasts for longer horizons.

Modelling students' experiences of learning statistics in a threshold concepts-enriched tutorial programme

Anisha Ananth[1] and Suriamurthee Maistry[2]

[1] *Durban University Of Technology*

[2] *University of KwaZulu-Natal, Durban, South Africa*

Corresponding author: anishas@dut.ac.za

Scholarship on the factors that affect students' learning in statistics have relied mainly on quantitative methodologies. As such, qualitative nuances as it relates to student learning remain relatively unexplored. To address this lacuna, this study applied a qualitative approach using a case study design to explore students' experiences of their learning in a threshold concepts-enriched statistics tutorial programme. Threshold concepts theory (Meyer & Land, 2003) and statistics pedagogy literature informed the tutorial programme activities. The larger part of the data was generated and initially analysed using Interactive Qualitative Analysis (IQA) which comprises two stages: focus groups and interviews. In the focus group phase, participants generated a view of learning on the programme at group level and the affinities (themes) identified by the focus group were arranged into a Systems Influence Diagram (SID) depicting the group's conception of their learning. The semi-structured individual interviews added depth to the focus group data as participants elaborated on their personal experiences with regard to each affinity. These findings reflect the duality of the cognitive and affective shifts students' experienced on their pedagogical pilgrimage - a metaphor used to describe the experiences and processes of students' learning in the threshold concepts-enriched tutorial programme. These findings are broadly consistent with the threshold concepts framework in highlighting that learning has strongly affective aspects entwined with the cognitive and once mastered have — transformative effects, and that disciplinary learning has implications for students' worldview and identity. The study has distinct implications for introductory statistics programme design and pedagogy.

A joint mixed model of adolescent's reproductive health service knowledge and utilization, and its associated factors in Jimma zone: A prospective longitudinal cohort Study

Tafere Tilahun Aniley[1], Legesse Kassa Debusho[1], Tadele Akeba Diriba[1] and Tefera Belachew Lema[2]

[1]*Department of Statistics, University of South Africa, Johannesburg*

[2]*Faculty of Public Health, Institute of Health, Jimma University, Ethiopia*

Corresponding author: 67146929@mylife.unisa.ac.za

Background: Adolescents who constitute one-third of the total population in Ethiopia, are usually exposed to reproductive health (RH) related problems. This is because of insufficient access to or inadequate knowledge of health services. Thus, the main aim of the current study was to investigate associated risk factors of adolescents' RH services knowledge and utilization in Jimma zone, Southwest of Ethiopia.Method: The data used in the study was taken from Jimma longitudinal family survey of youth study conducted in southwest Ethiopia. The responses measure adolescents' reproductive health service knowledge and utilization with binary outcomes. We proposed a bivariate logit mixed model to analyze both responses jointly, accounting for the correlation that exists within the data through random effects. Result: The results of the analysis with bivariate logit mixed model shows that the covariates gender, place of residence, current romantic relationship, and radio listening were significantly associated with both responses. However, adolescent age, society club participation, school attendance, and work status were significantly associated with adolescents' reproductive health services knowledge only. Whereas only current work status was significant covariate affecting adolescents' RH service utilization. Conclusion: Reproductive health service knowledge was not improved over the survey waves, while the exposure of adolescents to utilize RH service has increased. Based on the results we conclude that there was no clear evidence of early contact with adolescents to improve RH services knowledge. Finally, we recommend implementing various health intervention packages especially targetting adolescents to address the gap in knowledge and utilization of RH services.

A new generalized class of exponential ratio type estimators in ranked set sampling

Timothy Ayeleso[1], Olaniyi Olayiwola[2] and Abayomi Ajayi[2]

[1] Ogun State Bureau Of Statistics, Ogun State Civil Service, Abeokuta, Nigeria

[2] Department of Statistics, Federal University of Agriculture Abeokuta, Abeokuta, Nigeria

Corresponding author: ayelesobiyi@gmail.com

Ranked set sampling (RSS) is a good alternative for Simple Random Sampling for in sample survey. This study presents a generalized version of a class of exponential ratio type estimators in ranked set sampling (RSS) and compared with an existing generalized version of a class of modified exponential ratio estimators in simple random sampling (SRS). The data set used in this paper is the data on enrolment of students (variable of interest) and staff strength (auxiliary variable) in secondary schools in Egba zone of Ogun State Nigeria in 2015. The zone had 89 schools and a 3cycle ranked set sample of size 27 was selected. The descriptive statistics of the data set were obtained for the estimation of population ratio. The mean square errors (MSEs) for both the proposed estimators and modified SRS estimators were determined to obtain the efficiencies of the proposed estimators when g was set at 5, 2, 1, 0.75, 0.5 and 0.25. The population mean for student enrolment and staff strength were 1581.1 and 66.44 respectively which gave a population ratio of 23.80. At g equals 0.75, the MSEs of the modified SRS estimators were 168631.4, 164651.8, 155701.8, 165614.5, 167417.7, 163252.1, 155374.8, 151460.1, 165533.6 and 167714.7 while those of proposed estimators were 136330.3, 132447.5, 123715, 133386.8, 135146.1, 131081.8, 123396, 119576.4, 133307.8 and 135435.9 respectively. The MSEs for the members of the proposed generalized class of estimators were found to be smaller than those of SRS generalized class of estimators, hence they are more efficient estimators.

Some Dirichlet mixtures considered in a Bayesian context by using entropy for prior selection

Tanita Botha, Johannes T Ferreira and Andriette Bekker

Department of Statistics, Faculty of Natural and Agricultural Sciences,University of Pretoria, Pretoria, South Africa

Corresponding author: Tanita.Botha@up.ac.za

Random determinants play an essential role within multivariate analysis, but their distributions often present theoretical and computational challenges. To circumvent these challenges, this talk proposes a lower bound for the probabilistic analysis of the determinant emanating from a matrix consisting of independent but not necessarily identically distributed generalized beta entries. The 2×2 and 3×3 cases receive particular attention, and a brief simulation study verifies the results.

Clustering time-course data using P-splines and mixed effects mixture models.

Deidre Bredenkamp, Sollie Millard and Frans Kanfer

University Of Pretoria, Pretoria, South-Africa

Corresponding author: u04639864@tuks.co.za

This paper addresses cluster analysis of time-course data in a mixture model framework. To take into account the time dependency of such time-course data, as well as the degree of error present in many datasets, the mixed effects model with penalized B-splines has been presented. In this paper the performance of such a mixed effects model has been studied with regards to clustering of time course gene expression data in a mixture model system. The EM algorithm has been implemented to fit the mixture model in a mixed effects model structure. For each subject the best linear unbiased smooth estimate of its time course trajectory has been calculated and subjects with similar mean curves have been clustered in the same cluster. Model validation statistics such has the model accuracy and the coefficient of determination (R2) indicates that the model can cluster stochastic simulated data effectively into clusters that differ in either the form of the curves or the timings to the curves' peaks.The suggested technique is further evidenced by clustering time course gene expression data consisting of microarray samples from lung tissue of mice exposed to different Influenza strains from 14 timepoints.The results show a graphic overview of each cluster's genetic outcome, as well as the goodness-of-fit of the model via the 'mean curve' framework along with the respective confidence intervals.

Monitoring and mapping the critically endangered Clanwilliam cedar using aerial imagery and deep learning

Stefan Britz[1], Blessings Hadebe[1] and Glenn Moncrieff[2]

[1] *University Of Cape Town*

[2] *South African Environmental Observation Network*

Corresponding author: steviebritz@gmail.com

The critically endangered Clanwilliam cedar, *Widdringtonia wallichii*, is an iconic tree species endemic to the Cederberg mountains in the Fynbos Biome. Consistent declines in its populations have been noted across its range primarily due to the impact of fire and climate change. Mapping the occurrences of this species over its range is key to the monitoring of surviving individuals and is important for the management of biodiversity in the region. Recent efforts have focused on the use of freely available Google Earth^{TM} imagery to manually map the species across its global native distribution. This talk proposes an approach for automating the process of tree detection using deep-learning. The approach involves using sets of high-resolution red, green, blue (RGB) imagery to train artificial neural networks for the task of tree-crown detection. Additional models are trained on colour-infrared imagery, since live vegetation has a red tone on the near-infrared (NIR) spectrum. Preliminary results show that using an intersection-over-union threshold of 0.5 yields an average tree-crown recall of 0.59 with a precision of 0.46, and that the addition of the NIR spectral band does not result in improved performance. The viability of using this approach to regularly update maps of the Clanwilliam Cedar and monitor its population trends in the Cederberg is discussed.

Construction and analysis of experimental designs with three factors each having both fixed and random levels

Lyson Chaka

Sol Plaatje University, Kimberley, South Africa

Corresponding author: chakalyson@gmail.com

In experimental designs involving analysis of variance (ANOVA), the knowledge of whether effects are fixed, random or mixed is of paramount importance for the modelling process and interpretation of results. Classification of a factor as fixed or random effect depends on how a researcher selects the factor levels and the desired inference space. Increase of productivity and efficiency in the fourth industrial revolution era calls for development of new strategies and technologies that either replace or improve the old and existing ones. This leads to a shift in the way the fixed and random effects are conceptualised and structured as different factors. In this paper, we present the concepts and methods for designing and analysing experiments with three factors each consisting of both fixed and random levels. Consideration is made for two design structures namely, completely randomized design (CRD) and randomized complete block design (RCBD). The proposed approach allows for drawing of both narrow and broad inference for the same factor in a three-way treatment structure.

Failure rate monitoring in generalized gamma-distributed processes

Niladri Chakraborty and Tahir Mahmood

University Of The Free State, Bloemfontein, South Africa

Corresponding author: niladri.chakraborty30@gmail.com

With technological advancements, high-quality process monitoring has gained significant importance in the industry. Nowadays, most of the high-performing manufacturing processes produce a large number of conforming items with a few nonconforming items. Monitoring of time-between events is a well-known approach for real-time monitoring of these highly efficient processes. Usually, it is assumed that the time-between-events follow an exponential or gamma distribution. However, the generalized gamma distribution is one of the most popular choices for modeling skewed data. Monitoring of skewed processes poses a challenge in designing an unbiased monitoring scheme where the probability of signal should be higher than the size. In this work, we consider a two-sided monitoring scheme based on the generalized gamma distribution. This would provide a one-stop solution to the two-sided monitoring of many skewed distributions. We also proposed a generalized analytical solution to the unbiased design of a skewed monitoring scheme. The extensive numerical study showed encouraging performance properties. A couple of practical applications in connection to monitoring renewable energy and coal mine explosions have been discussed.

Modelling the output from a commercial chemical facility using the Cox proportional hazard regression.

Roelof Coetzer[1], Jaco Visagie[1], Marius Smuts[1], James Allison[1] and Alta de Waal[2]

[1] *North-west University, Vanderbijlpark, South Africa*

[2] *University of the Free State, Bloemfontein, South Africa*

Corresponding author: roelof.coetzer@nwu.ac.za

The Cox proportional hazard (CPH) regression model has been used with great success for modelling the time until certain events occur, and for studying the dependency of survival time or time until event on predictor variables. In this paper, we illustrate that the CPH model can also be used for modelling a response variable, which is heterogeneous over time, as a function of predictor variables. In addition, we evaluate a number of alternative distributions for the baseline hazard and quantify the accuracy of the CPH model in predicting the response variable using the predicted root mean square error. The accelerated failure time (AFT) model, where the explanatory variables act multiplicatively on time, is considered as an alternative to the CPH model. The CPH and AFT models are used for the prediction of a key process variable in a commercial chemical facility.

Using Principled Bayesian inference to assess the viability of wind power in South Africa

Matthew De Bie

University Of Pretoria, Johannesburg, South Africa

Corresponding author: u18115943@tuks.co.za

In light of the ongoing load-shedding crisis, South Africa must look towards sources of renewable energy to supplement its ageing, coal-reliant power infrastructure. Sites in the coastal regions of South Africa are locations where exploitation of wind energy may be feasible. To begin our investigation, we will fit South African wind speed data to a Weibull model. Existing literature disagrees about which estimation method is best suited to estimate the Weibull shape parameter. Through simulation, we will contrast several estimation methods and recognise the Bayesian application of the PC Prior as most appropriate for estimating the shape parameter of our proposed Weibull model. We will then apply this Bayesian framework, through several models of increasing complexity, to actual wind speed data by means of the R-INLA package. We aim to gain a better understanding of the functional relationship between recorded wind speeds and the altitude, temporal and spatial conditions under which these measurements were taken. Ultimately, our goal is to construct a statistical framework through which the feasibility of harnessing wind energy in these coastal regions may be evaluated.

Seasonal and station effects modelling to extreme temperature data in South Africa

Legesse Kassa Debusho and Tadele Akeba Diriba

University Of South Africa, Florida, Johannesburg, South Africa

Corresponding author: debuslk@unisa.ac.za

Background: Adolescents who constitute one-third of the total population in Ethiopia, are usually exposed to reproductive health (RH) related problems. This is because of insufficient access to or inadequate knowledge of health services. Thus, the main aim of the current study was to investigate associated risk factors of adolescents' RH services knowledge and utilization in Jimma zone, Southwest of Ethiopia.Method: The data used in the study was taken from Jimma longitudinal family survey of youth study conducted in southwest Ethiopia. The responses measure adolescents' reproductive health service knowledge and utilization with binary outcomes. We proposed a bivariate logit mixed model to analyze both responses jointly, accounting for the correlation that exists within the data through random effects.Result: The results of the analysis with bivariate logit mixed model shows that the covariates gender, place of residence, current romantic relationship, and radio listening were significantly associated with both responses. However, adolescent age, society club participation, school attendance, and work status were significantly associated with adolescents' reproductive health services knowledge only. Whereas only current work status was significant covariate affecting adolescents' RH service utilization.Conclusion: Reproductive health service knowledge was not improved over the survey waves, while the exposure of adolescents to utilize RH service has increased. Based on the results we conclude that there was no clear evidence of early contact with adolescents to improve RH services knowledge. Finally, we recommend implementing various health intervention packages especially targetting adolescents to address the gap in knowledge and utilization of RH services.

A spatially explicit modelling strategy for Covid-19 predictions and 4th wave risk analysis in Gauteng

Claudia Dresselhaus[1], Inger Fabris-rotelli[1], Raeesa Manjoo-Docrat[2], Nada Abdelatif[3], Nontenbeko Dudeni-Tlhone[4], Pravesh Debba[4], Jenny Holloway[4], Charl Janse van Rensburg[3,4], Renate Thiede[1] and Sibusiswe Makhanya[5]

[1] *University Of Pretoria*

[2] *University of the Witwatersrand*

[3] *Medical Research Council*

[4] *CSIR*

[5] *IBM Research*

Corresponding author: inger.fabris-rotelli@up.ac.za

The COVID-19 mass vaccination roll-out plan is designed to allocate vaccinations according to a three-phase strategy that prioritises frontline healthcare workers and the elderly, especially those who are most likely to present with comorbidities. Current studies show overwhelming evidence that vaccinations protect people from severe COVID-19 outcomes. Given this context, the public and policymakers may already be concerned about the timing and severity of the fourth wave of the pandemic locally. Given over 18 months of data on infection outcomes, recent data on vaccination rates and vaccination centre locations, as well as more clarity and data on risk and vulnerability factors, there is imperative to consider nuanced Susceptible-Exposed-Infectious-Removed (SEIR) models to predict the expected number people with severe outcomes from infection in the fourth wave of the pandemic. In our work, we aim to incorporate the vaccinated individuals into a spatial SEIR model to form a SEIRV model. This model will be used to investigate future waves of the COVID pandemic given that the vaccination roll-out in South Africa.

A Spatial SEIR Model for COVID-19 in South Africa

Inger Fabris-rotelli[1], Jenny Holloway[2], Zaid Kimmie[3], Sally Archibald[4], Pravesh Debba[2], Raeesa Manjoo-Docrat[4], Alize le Roux[2], Nontembeko Dudeni-Tlhone[2], Charl Janse van Rensburg[5], Renate Thiede[1], Nada Abelatif[4] and Arminn Potgieter[1]

[1] *University Of Pretoria*

[2] *Council for Scientific Research*

[3] *Foundation for Human Rights*

[4] *University of the Witwatersrand*

[5] *South African Medical Research Council*

Corresponding author: inger.fabris-rotelli@up.ac.za

The virus SARS-CoV-2 has resulted in numerous modelling approaches arising rapidly to understand the spread of the disease COVID-19 and to plan for future interventions. Herein, we present an SEIR model with a spatial spread component as well as four infectious compartments to account for the variety of symptom levels and transmission rate. The model takes into account the pattern of spatial vulnerability in South Africa through a vulnerability index that is based on socioeconomic and health susceptibility characteristics. Another spatially relevant factor in this context is level of mobility throughout. The thesis of this study is that without the contextual spatial spread modelling, the heterogeneity in COVID-19 prevalence in the South African setting would not be captured. The model is illustrated on South African COVID-19 case counts and hospitalisations.

Entry-level statistics supervisor development in South Africa

Inger Fabris-rotelli[1], Sonali Das[1], Ansie Smit[1], Gao Maribe[1], Michael von[2], Danielle Roberts[3], Fabio Correa[4] and Daniel Maposa[5]

[1] *University Of Pretoria, South Africa*

[2] *University of the Free State, South Africa*

[3] *University of Kwazulu-Natal, South Africa,*

[4] *Rhodes University, South Africa*

[5] *University of Limpopo, South Africa*

Corresponding author: inger.fabris-rotelli@up.ac.za

In 2020 a group of 8 novice, and near-novice, doctoral supervisors in academic Statistical sciences in South Africa initiated the use of the portfolio developed under this project with their new and current doctoral students. Biannual meetings, as well as virtual meetings every week, have documented the feedback, hurdles and successes of the portfolio. We present our discussions and suggestions for future work in this talk, including publications, research dissemination at conferences and similar, as well as continued mentorship.

Automatic Generation of Online Statistics Assessments Using R-exams

Thomas Farrar

Cape Peninsula University Of Technology, Cape Town, South Africa & University of the Western Cape, Cape Town, South Africa

Corresponding author: farrart@cput.ac.za

Online modes of assessment in higher education have been pushed to the forefront by the COVID-19 pandemic. Online tests have certain advantages over "pen-and-paper" tests in areas such as authenticity and "assessment for learning," but also bring challenges in areas such as integrity. e-Learning platforms (LMSs) have considerable functionality for creating online tests but also significant limitations. For instance, there is limited functionality for randomisation of questions (an important bulwark for integrity). Furthermore, it is labour-intensive to add rich content (tables, figures, mathematical expressions) and manually produce memos within LMS interfaces. The `exams` package in R statistical software harnesses the computing power of R to generate versatile assessments that can then be imported into the LMS for deployment. This presentation will provide an overview of the functionality of the exams package and the implementation workflow. Examples will be provided of different assessment tasks that can be automated, including not only data analysis and visualisation tasks but also mathematical statistics questions where the objective is a mathematical expression or a proof. The presentation will focus primarily on Blackboard, the LMS software used by the author's institution. However, to make the presentation as widely accessible as possible, implementation in all other LMS software used at South African universities (Moodle, Sakai, Canvas, and Brightspace) will also be discussed. Results of a student feedback survey regarding student experience of these assessments will be shared.

A New Approach to Error Variance Estimation in a Heteroskedastic Linear Model

Thomas Farrar[1,2], Retha Luus[2] and Renette Blignaut[2]

[1] *Cape Peninsula University Of Technology, Cape Town, South Africa*

[2] *University of the Western Cape, Cape Town, South Africa*

Corresponding author: farrart@cput.ac.za

Estimation of error variances is an important precursor to both estimation of and inference on regression coefficients in a heteroskedastic linear model. The variance estimates can be used to compute heteroskedasticity-consistent standard errors for coefficient estimates as well as to compute weights for feasible weighted least squares estimation. Most existing heteroskedasticity-consistent covariance matrix estimator (HCCME) methods make element-wise bias corrections to the squared ordinary least squares (OLS) residuals, e_i^2. The corrections are however based on the conditional expectation of the e_i^2 under homoskedasticity, not under heteroskedasticity. A proposed new approach treats the conditional expectation of the e_i^2 under heteroskedasticity as the conditional mean function of an auxiliary regression model with the e_i^2 as the responses. Two methods for reducing the dimensionality of the resulting parameter space are considered. The first method, HC8, assumes a functional relationship between the error variance parameters and the design variables from the main model. The second method, HC9, assumes that certain subsets of the observations have equal error variances based on their proximity in the design space. Appropriate subsets may be computed using agglomerative hierarchical clustering. In either case, the auxiliary regression model is nonlinear and is fit using a quasi-likelihood procedure. A simulation experiment is performed to compare the new methods to existing HCCMEs. The problem of feature selection in the auxiliary model is discussed.

Some simple statistical ideas and techniques useful for understanding the COVID-19 pandemic

Paul Fatti

Wits University, Johannesburg, South Africa

Corresponding author: paulfatti@gmail.com

The presentation will discuss the following topics relating to the pandemic: 1,Estimating the number of infections. 2, Estimating the number of deaths. Screening for COVID-19. 3,Optimal group testing for COVID-19 by pathology laboratories. 4,Testing a vaccine .5, Each of these topics will be discussed, using generally simple statistical concepts and techniques, giving interesting insights and some surprising results.

An insight to inference on the probabilistic determinant of independent generalized beta entries

Johan Ferreira, Andriette Bekker, Tanita Botha and Seite Makgai

University Of Pretoria, Pretoria, South Africa

Corresponding author: johan.ferreira@up.ac.za

Random determinants play an essential role within multivariate analysis, but their distributions often present theoretical and computational challenges. To circumvent these challenges, this talk proposes a lower bound for the probabilistic analysis of the determinant emanating from a matrix consisting of independent but not necessarily identically distributed generalized beta entries. The 2×2 and 3×3 cases receive particular attention, and a brief simulation study verifies the results.

Bayesian Spatial Model for Disease Mapping: Application to HIV Distribution in Ethiopia

Leta Lencha Gemechu and Legesse Kassa Debusho

University of South Africa, Johannesburg, South Africa

Corresponding author: 64875636@mylife.unisa.ac.za

In this study, we applied a Bayesian spatial model to investigate the spatial distribution of HIV in Ethiopia, using district level aggregated HIV cases obtained from the Demographic and Health Survey data (EDHS, 2016). Both informative and non-informative priors were used. The informative priors for coefficients were formulated from previous EDHS data. Whereas the prior for spatial random effect is defined in framework of penalized complex prior, where the pooled spatial variance obtained from previous two surveys used as upper bound or limit of sampling variance. Our proposed model was compared to commonly used models using observed and simulated data. Among the models considered, mofified Besag, York and Mollie (BYM2) model (with informative prior) had fitted the data best, therefore, investigation of factors affecting spatial association of HIV prevalence examined based on this model. The results show the likelihood of being infected by HIV virus varies across clusters and regional location. For instance, High clusters of HIV cases were observed in Gambella region, Harari, Addis Ababa, and borderline districts located in Tigray and Amhara regional states. In addition to regional variation, significant difference was seen with respect to place of residence (Urban or Rural) and gender, individuals in urban areas and women highly affected by HIV burden, with ratios of about 7 to1 and 6 to 1 (compared to rural dwellers and men) respectively. Overall, our study result revealed strong spatial disparity of HIV distribution at different geographical levels. Furthermore, women's living in urban areas are the top affected social group.

Binary Particle Swarm Optimization(BPSO) based feature selection

Michelle Gilfillan, Sollie Millard and Frans Kanfer

University of Pretoria

Corresponding author: u16095503@tuks.co.za

This paper studies feature selection using Binary Particle Swarm Optimization(BPSO) for high dimensional data sets. Logistic regression and k-nearest neighbour(KNN) classifiers are used. BPSO based feature selection uses a meta-heuristic search strategy to find near optimal feature subsets in a small amount of time. These methods are compared with the results of a random forest classifier. Theoretical aspects together with an application are proposed.

Robust joint modelling of longitudinal data and survival data: detection and downweighting of longitudinal measurements

Freedom Gumedze

Department of Statistical Sciences, University of Cape Town

Corresponding author: Freedom.Gumedze@uct.ac.za

Mixed-effects location scale models allow simultaneous modelling of between-subject and within-subject variability. These models include log-linear models for the between-subject and within-subject variability. The log-linear models could potentially include covariates. The models assume that the residual errors and the random effects are normally distributed. This makes them sensitive to outliers. These models have been extended to joint models of longitudinal data and time-to-event data. We explore Cook-type influence diagnostics for the mixed-effects location scale model, assumed for the longitudinal sub-model, and an approach to down-weight outlying subjects. We illustrate the methods using data from a large cardiology clinical trial.

Dynamic prediction of virologic failure in a cohort of HIV infected individuals on antiretroviral therapy in the Western Cape

Frissiano Honwana, Elton Mukonda, Freedom Gumedze, Marvin Hsiao, Landon Myer and Maia Lesosky

University of Cape Town, Cape Town, South Africa

Corresponding author: honwana@ymail.com

Background: Personalized medicine is receiving greater attention as health data collection rapidly digitizes and methodological development has seen an increase in statistical models for individual prediction. We consider dynamic prediction models to model the dependency between longitudinal viral load (VL) and virologic failure (VF) in people living with HIV. Methods: We included 91,818 individuals with longitudinal VL measures from routine data between 2008 and 2018. We used a shared random effects model (SREM) to predict virologic failure based on historical VL trajectory and baseline characteristics. Time-dependent area under the curve (AUC) of the receiver operating characteristics (ROC) curve was used to quantify the prediction accuracy. Results: The SREM fit was acceptable with residual diagnostics satisfying the assumptions of the model. The SREM demonstrated good prediction accuracy with AUCs ranging from 0.69 to 0.76 Conclusion: SREM effectively incorporates baseline covariate with time-varying viral load and continuously updates virological failure probabilities with every new additional repeated measurement. The dynamic predictions from the SREM using routinely captured data provides an opportunity to flag individuals who may be at greater risk for negative outcomes. This may be a first step in individualized care models for people living with HIV.

Data-driven methods for subgroup identification in clinical trials with an application

Charl Janse van Rensburg, Din Chen and Samuel Manda

Biostatistics Research Unit, South African Medical Research Council, Pretoria

Corresponding author: charl.jansevanrensburg@mrc.ac.za

Randomized clinical trials provides the best evidence on the efficacy of new therapeutic drugs. In many trials, the main treatment effect of interest may not be significant. Post hoc analysis may be conducted to identify subgroups for whom the treatment may have worked. However simple methods suffer from low power, as well as bias. In the last two decades many data-driven approaches have been developed to identify subgroups of patients with treatment effect in failed trials, or enhanced treatment effect compared to the overall effect. Subgroup treatment effect identification could be approached by using decision trees, random forests, support vector machines, as well as model-based approaches. We introduce some of these methods, most notably the GUIDE algorithm, and compare the methods under a simulation study. The methods are also applied to a real-life trial data set. Recommendations are made for good practice when doing exploratory subgroup analysis using these methods.

Estimation of Complier Average Causal Effect using Proportional Hazards Models in Randomised Trials with Competing Risks

Andreas Kryger Jensen

University of Copenhagen

Corresponding author: aeje@sund.ku.dk

Randomised studies are ideal to learn about causal effects. A common problem in such studies, however, is that all subjects may not comply to the treatment they are randomized to receive. It is well-known that the as-treated analysis my render incorrect results and that the intention-to-treat analysis is not targeting the causal effect of the treatment. However, instrumental variable techniques can be used to estimate the causal effect. We consider the situation with a survival outcome using a structural proportional hazards model for the compliers under the active treatment. Subjects under the control treatment do not have access to the active treatment. We present an estimator for the complier average causal effect (CACE) in the proportional hazards setting and give its large sample properties. We also illustrate the extension of this problem to the setting of a competing risks outcome.

Probabilistic Graphical Models and Belief Propagation: Approximations via Free Energies

Francois Kamper

Stellenbosch University, Stellenbosch, South Africa

Corresponding author: francoisk@sun.ac.za

A probabilistic graphical model (PGM) aims to illustrate dependencies between random variables arising from multivariate distributions. These dependencies can depict dependencies such as causality (Bayesian networks) and conditional independence (Markov networks). Given a PGM, and associated graph structure, one is typically tasked with performing inference on the underlying multivariate distribution. The inference task typically involves tasks such as determining multiple marginal distributions or determining the maximum a posteriori (MAP) assignment. A key reason for the popularity of PGMs is the existence of algorithms that can exploit sparsity in the graph structure to perform inference in a computationally efficient way. Inference can be done either exactly or approximately, where the latter focuses on computational speed rather than inference accuracy. Perhaps one of the more famous approximate inference algorithms is called belief propagation (BP), where a Markov network (say) is converted to another graph type (such as factor or cluster graph) on which inference is then performed. BP is an example of a message-passing algorithm, where nodes in the graph perform operations in parallel, and then communicate the results to neighboring nodes. The purpose of this talk is to introduce belief propagation as a solution to an optimization problem, where a distribution is approximated by a specific type of factorization inspired by tree-structured graph typologies.

Handling incomplete data using random draws

Nyiko Muhluri Khoza

University of South Africa

Corresponding author: ngobenh@unisa.ac.za

The study observes gaps in the current methods that handle the missing observations problem in the data. The gaps identified are also addressed by evaluating variable importance using the HPSPLIT procedure and the study compares the samples before and after data cleaning to show the effectiveness of using the HPSPLIT procedure. We argue that each missing observation has a random chance of occurrence as with the observed observations, however, this study considers value formats of the data when handling the incomplete problems in the data. The study proposes a method that does not violate value formats of the data when compared to the current methods that handle the incomplete data. The proposed method retains the sample size and give biased estimates when comparing the handled to the selected sample to estimate the true population. The current techniques that uses imputation to handle missing observations problems in data has shown through this study that not all imputation techniques are good in handling the incomplete data problem. The study conducts experiments on the current methods of handling the incomplete data problems and assesses the imputation techniques' reliability. The focus is on correcting the gaps identified in each method as shown in the experiments when looking at each of the current techniques evaluated by the study. The proposed method is a technique that multiplly imputes random draws of observations until convergence is reached, focusing on the value formats of the data. The proposed method also cleans the imputed observations to fix or to resolve the random–draw method shortfall.

On the use of Voronoi tessellations for detection of spatial inhomogeneity in regular spatial point patterns

Christine Kraamwinkel, Inger Fabris-Rotelli, Rian Botes, Kabelo Mahloromela and Ding-Geng Chen

University Of Pretoria, Pretoria, South Africa

Corresponding author: christine.kraamwinkel@up.ac.za

Much spatial analysis requires the division of the spatial window into equal sized quadrats. Specifically, tests for homogeneity of spatial point patterns use the counts of points in each quadrat to determine the homogeneity. A choice has to be made on the quadrat size, thereby introducing a hyperparameter that must be chosen appropriately. In this paper, we instead partition the spatial window using Voronoi tessellations. A Voronoi tessellation is the partition of the spatial window into convex polygons, called Voronoi cells, consisting of all points of the plane closer to that point than to any other. We show that the shape measures of the polygons can differentiate between a homogeneous and inhomogeneous regular spatial point pattern. Four measures of elongation for Voronoi cells were investigated, namely circularity, radial radius, aspect ratio and elongation, through a simulation study and real data application. The results indicate future development of a robust hypothesis test for homogeneity is possible.

Mediation analyses with survival outcomes

Theis Lange

University of Copenhagen

Corresponding author: thlan@sund.ku.dk

In this talk I will present how causal inference methods allow us to rigorously decompose the effect of an exposure on a survival outcome. I will also present how natural effect models allow us to estimate this even with a (semi) high dimensional mediator. Finally, I will provide suggestions for future research.

Age and Size related Reference Ranges for lung function measurements of a cohort of South African children

Francesca Little, Carlyle McCready, Diane Gray, Shaakira Chaya, Heather Zar and Lesley Workman

University Of Cape Town

Corresponding author: francesca.little@uct.ac.za

The use of growth percentiles for anthropometric measurements to monitor childhood development is well known. Growth reference standards are derived based on a large representative multi-country and multi-ethnical cohort of children from birth. Childhood growth is then monitored by either comparing actual growth to the percentiles of these growth standards or by calculating a “z-score” that measures the deviation from “normal” growth. The derivation of the centiles and z-scored are based on the modelling of the moments of the underlying distribution of the growth measurements, the mean (or median), standard deviation (or coefficient of variability), the skewness and the kurtosis using cubic or basis splines to capture the nonlinear association with age. The most common methodology for doing this is the technique known as generalized additive models for location, scale and shape (GAMLSS) that extends and incorporates the much used LMS method. Reference ranges are not only important for anthropometric growth but also for other medical measurements, for example laboratory reference ranges and lung function measurements. These measurements often depend not only on age but also on size (for example, height of children), and hence the construction of reference ranges need to take size into account. The GAMLSS methodology allows for a relatively easy incorporation of size in the modelling of the moments of the outcome distributions either as additive or multiplicative factors. We illustrate the derivation of reference ranges for lung function measurements in a cohort of South African children from 6 weeks to 5 years of age.

Markov chains, detailed balance and time-reversibility

Iain MacDonald and Etienne Pienaar

University Of Cape Town

Corresponding author: iain.macdonald@uct.ac.za

The use of growth percentiles for anthropometric measurements to monitor childhood development is well known. Growth reference standards are derived based on a large representative multi-country and multi-ethnical cohort of children from birth. Childhood growth is then monitored by either comparing actual growth to the percentiles of these growth standards or by calculating a “z-score” that measures the deviation from “normal” growth. The derivation of the centiles and z-scored are based on the modelling of the moments of the underlying distribution of the growth measurements, the mean (or median), standard deviation (or coefficient of variability), the skewness and the kurtosis using cubic or basis splines to capture the nonlinear association with age. The most common methodology for doing this is the technique known as generalized additive models for location, scale and shape (GAMLSS) that extends and incorporates the much used LMS method. Reference ranges are not only important for anthropometric growth but also for other medical measurements, for example laboratory reference ranges and lung function measurements. These measurements often depend not only on age but also on size (for example, height of children), and hence the construction of reference ranges need to take size into account. The GAMLSS methodology allows for a relatively easy incorporation of size in the modelling of the moments of the outcome distributions either as additive or multiplicative factors. We illustrate the derivation of reference ranges for lung function measurements in a cohort of South African children from 6 weeks to 5 years of age.

Panel Data Changepoint Estimation via Regularization

Matúš Maciak

Charles University, Prague, Czech Republic

Corresponding author: maciak@karlin.mff.cuni.cz

Implied volatility (IV) is used as a general but powerful tool for analyzing financial markets. We propose a novel approach to estimate the overall IV dynamics represented by an underlying panel data model with changepoints. A robust semi-parametric regression framework and atomic pursuit techniques lasso based regularization in particular are applied to estimate the underlying analytical structure of the IV surface and a formal statistical test is used to detect significant changepoints. The overall complexity of the model assumes changepoints which may occur over time, in the analytical structure of the IV smiles, or both. Theoretical and practical details are discussed and the main statistical properties are derived. Empirical properties are investigated in a simulation study and real-life applications are presented to illustration wide and general applicability.

Covariate construction of nonconvex windows for spatial point patterns

Kabelo Mahloromela, Inger Fabris-Rotelli and Christine Kraamwinkel

University Of Pretoria, Pretoria, South Africa

Corresponding author: u14194237@tuks.co.za

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

Predictive modelling of Covid-19 new cases and deaths in South Africa

Kajingulu Malandala

University Of South Africa, South Africa

Corresponding author: malank@unisa.ac.za

Coronavirus pandemic had already affected more than 8.4 million people in the African continent leading to an estimated 214000 deaths. South Africa (SA) is one of the most affected country in the continent with the highest number of cases. The literature related to Covid-19 is limited and has been focusing on modelling and prediction of the disease in the early stages of the outbreak. In the current study, we propose Generalized Additive model for location, scale and shape models to predict the number of Covid-19 cases and fatalities in SA. Generalized Additive models for location, scale and shape (GAMLSS) models are extension of the generalized linear models (GLM) and the generalized additive models (GAM) with location, scale and shape parameters which are modelled as linear, nonlinear or smooth function of the covariates. The results suggest that GAMLSS approach is flexible and allow us to produce reliable estimate of the variance at each point of time and the distribution of expected values in the future.

Monitoring multivariate profiles using the quadruple exponentially weighted moving average scheme with fixed and random exploratory variables

Jean-claude Malela-Majika

University Of Pretoria, Pretoria, South Africa

Corresponding author: malela.mjc@up.ac.za

When the quality process is characterised by a functional relationship between a dependent variable and one or several explanatory variables, classical monitoring schemes become inappropriate and unresponsive. In this case, profile or (regression) monitoring schemes are recommended. This paper proposes new quadruple EWMA (QEWMA) schemes for monitoring linear profile data using fixed and random exploratory variables. In zero-state, the proposed schemes are found to be more responsive for a large range of shifts in the regression parameters and error variance, while in steady-state, the EWMA scheme is more responsive to different shifts as compared to other memory-type schemes considered in this study. Real-life data are used to demonstrate the application and implementation of the newly proposed schemes.

Comparing Effect of HIV Treatment Regimens on Time to Mortality, and Virological Failure and Rebound among HIV Positive Patients using Inverse Probability of Treatment Weighting Estimation of Marginal Structural Models

Samuel Manda[1] and Halima Twabi[2]

[1] *Biostatistics Research Unit, South African Medical Research Council, Pretoria*

[2] *Department of Mathematical sciences, University of Malawi, Zomba, Malawi*

Corresponding author: samuel.manda@mrc.ac.za

Text missing

Correlated gamma frailty models for bivariate survival time data

Adelino Martins

Eduardo Modlane University, Maputo, Mozambique

Corresponding author: adelio.martins33@gmail.com

Frailty models have been developed to quantify both heterogeneity as well as association in multivariate time-to-event data. In recent years, numerous shared and correlated frailty models have been proposed in the survival literature allowing for different association structures and frailty distributions. A bivariate correlated gamma frailty model with an additive decomposition of the frailty variables into a sum of independent gamma components was introduced before. Although this model has a very convenient closed-form representation for the bivariate survival function, the correlation among event- or subject-specific frailties is bounded above which becomes a severe limitation when the values of the two frailty variances differ substantially. In this paper, we review existing correlated gamma frailty models and propose novel ones based on bivariate gamma frailty distributions. Such models are found to be useful for the analysis of bivariate survival time data regardless of the censoring type involved. The frailty methodology was applied to right censored and left-truncated Danish twins mortality data and serological survey current status data on varicella-zoster virus and parvovirus B19 infections in Belgium. From our analyses, it has been shown that fitting more flexible correlated gamma frailty models in terms of the imposed association and correlation structure outperforms existing frailty models.

Bayesian quantile regression analysis for stroke predictors in South Africa using MCMC method

Lyness Matizirofa[1] and Delson Chikobvu[2]

[1] *University of South Africa, Johannesburg, South Africa*

[2] *University of the Free State, Bloemfontein, South Africa*

Corresponding author: lmatizi@yahoo.com

Background: In South Africa (SA), stroke is the second highest cause of mortality and disability. Yet, little is known about the modelling modifiable and non-modifiable stroke predictors. Bayesian quantile regression (BQR) can be used for this type of modelling. This paper provides a quantile inference approach through the Bayesian modelling approach. Identification of stroke predictors using appropriate statistical methods can help formulate appropriate health programs and policies aimed at reducing the stroke burden. Analysis of stroke predictors have in the main, concentrated on mean regression, yet modelling with quantile regression (QR) is more appropriate than using mean regression. This is because the QR provides flexibility to analyse the stroke predictors corresponding to quantiles of interest. This study aims to identify and quantify stroke predictors, through BQR analysis. Methods: Hospital-based data from 35 730 stroke cases were retrieved from selected private and public hospitals between January 2014 and December 2018. The Markov chain Monte Carlo (MCMC) method is used for obtaining posterior distributions of the parameters of interest. The Bayesian approach is compared to the classical approach. Results: Of the 35730 stroke cases, 22183 were diabetic. The age groups 55-75 and 76-98 years, female gender and black race had a bigger effect on stroke distribution at the lower than upper quantiles. Diabetes, cholesterol, heart problems and hypertension showed a significant impact on stroke distribution ($p < 0.0001$). Conclusions: Modelling stroke predictors using BQR can provide information beneficial for addressing the stroke burden in SA.

Modelling the spread of COVID-19 in South Africa using stratified compartmental models in the period March 2020 - August 2020

Elona Mbayise[1], Raeesa Docrat[1], Nada Abelatif[2], Pravesh Debba[3], Inger Fabris-Rotelli[4], Jenny Holloway[3], Nontembeko Dudeni-Tlhone[3], Renate Thiede[4], Charl Janse van Rensburg[2] and Sibusiswe Makhanya[5]

[1] *University of the Witwatersrand*

[2] *Medical Research Council*

[3] *Council for Scientific and Industrial Research*

[4] *University of Pretoria*

[5] *IBM Research*

Corresponding author: inger.fabris-rotelli@up.ac.za

The novel coronavirus strand (SARS-CoV-2) first appeared in Wuhan, China in December 2019 and caused the respiratory syndrome COVID-19. A unique feature of COVID-19 is its non-uniform effect on populations. The effects of COVID-19 are more severe amongst the older and people with co-morbidities as seen by the higher mortality, infection and hospitalisation rates observed amongst these groups. This study models the spread of COVID-19 in South Africa March-August 2020 using stratified compartmental models to capture the population heterogeneity. An age and co-morbidity stratified compartmental model was built with additional compartments to capture the unique dynamics of COVID-19. A sensitivity analysis was performed to determine the models' sensitivity to start date and lockdown level to determine the optimal start date and to identify the effects of harsh lockdown restrictions on infections and hospitalisations. A parameter sensitivity analysis was also conducted to determine the parameters that needed to be re-estimated to improve model accuracy and to identify the age groups which were driving infections, hospitalisations, and deaths. These analyses showed that a prolonged harsh lockdown would have reduced infections by approximately 50% and delayed the infection peak by approximately 4 months. The analyses also showed that hospitalisations were driven by the 61-75 age group while infections and deaths were driven by the 76-90 age group. In addition, the model was most sensitive to infection duration, death rate and proportion of asymptomatic infection. These parameters were re-estimated to better capture the age and co-morbidity dependent dynamics of COVID-19.

A comparative study of the stylized facts of South African and Indian Stock markets

Lindokuhle Mbhense, Sen Riturpana and Syamala Krishnannair

University of Zululand

Corresponding author: lindokuhlesandile34@gmail.com

The study looks at different stylized facts within the South African market (JSE Top 40 Index)'s historical data from the finance.Yahoo.com was employed for analysis. A closer look at the behaviour of the South African market brings lot of interest since this is one of the most developing markets, which then gives the perfect opportunity for researchers to develop reliable models for forecasting returns and hence price derivation. It was obtained that most of the stocks in JSE Top 40 Index showed larger upward movements than draw-downs similar to Indian stock market, which makes South African market a promising market to invest in. Some of the stocks revealed the presence of auto-correlation which can be a tool for predicting future prices which favors the investors. The behavior of the South African market is almost the same as that of the Indian stock market regarding stylized facts.

Wheezing Phenotypes And Early-Life Determinants In A South African Birth Cohort Study.

Carlyle McCready[1], Sadia Haider[2], Francesca Little[1], Lesley Workman[3], Diane Gray[3], Mark Nicol[4], Adnan Custovic[2] and Heather J. Zar[3]

[1] *Department of Statistical Science, University of Cape Town, South Africa*

[2] *National Heart and Lung Institute, Imperial College London, UK*

[3] *SA-MRC Unit on Child & Adolescent Health, and Dept Paediatrics & Child health, University of Cape Town, South Africa*

[4] *Infection and Immunity, School of Biomedical Sciences, University of Western Australia, , Australia*

Corresponding author: MCCCAR007@myuct.ac.za

Objective: We aimed to identify underlying latent patterns of wheezing, and associated risk factors, in a South African birth cohort. Methods: Wheezing was longitudinally identified from birth to 5 years for infants from the Drakenstein Child Health Study. Using repeated binary indicators denoting the presence/absence of wheeze, we derived a set of multi-dimensional indicators which incorporates a spells approach to describe the temporal characteristics of wheeze. These indicators were clustered using the Wishart distance matrix and partitioning around medoids (PAM) algorithm to identify homogenous phenotypes of wheeze. Multinomial logistic regression models were used to investigate phenotype specific risk factors. The stability and validity of the underlying latent phenotypes were investigated using a repeated sampling approach. Results: Wheezing was common in 455/950 (48%) children. Four phenotypes were identified: never-wheezing (495, 52%), early-transient wheezing (202, 21%), late-onset wheezing after 1 year (104, 11%), and recurrent wheeze (149, 16%). Early-life lower respiratory tract infection (LRTI) was a strong risk factor associated with all wheezing phenotypes, but most strongly with recurrent wheeze, as was the concomitant presence of the respiratory syncytial virus (RSV), rhino and adeno viruses, which are viral pathogens known to cause respiratory infections in children. Other factors associated with recurrent wheeze were maternal smoking, intimate partner violence, higher socioeconomic class, or male child. Conclusion: Childhood wheezing represents a heterogenous airway disease with specific identifiable phenotypes and associated risk factors in African children. Early life LRTI and environmental factors influence wheezing risk.

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

Nobuhle Mchunu

South African Medical Research Council (SAMRC), Durban, South Africa, & University of KwaZulu-Natal (UKZN), Pietermaritzburg, South Africa,Centre for the AIDS Programme of Research in South Africa (CAPRISA), Durban, South Africa

Corresponding author: nobuhle.mchunu@mrc.ac.za

Background: Joint modeling is the most appropriate method for studying potential associations between biomarkers and time to event outcomes. The association structure linking the two sub-models is of fundamental importance in the joint modeling framework. However, rationale for selecting this association structure has received little attention in the literature. To this end, we aim to explore five alternative association structures between the CD4 count and the risk of death and ultimately select the best association structure for our data. Methods: We used data from CAPRISA, the Starting Antiretroviral Therapy at Three Points in Tuberculosis (SAPIT) study, an open-label, three armed randomised, controlled trial between June 2005 and July 2010 (N=642). We used the Deviance Information Criterion (DIC) to select the final model, with smaller values indicating better model adjustments to the data. Results: Among the 642 patients enrolled in the SAPIT trial, 214 (33.3%) were in the early integrated arm, 215 (33.5%) in the late integrated arm and 213 (33.2%) in the sequential arm. Patient characteristics were similar across the three study arms. The joint model with random effects was chosen as our best model where the baseline levels of the underlying square root CD4 count as well as the longitudinal evolution of the CD4 count were found to be strongly related to the hazard of death. Conclusions: The current value association structure may not always be appropriate in expressing the correct association between the outcomes in all settings. Thus, exploring other clinically meaningful association structures linking the two processes expands the usefulness of the joint modeling framework.

Big data analytics through ridge-type and Liu-type estimators

Salomi Millard[1], Mohammad Arashi[2] and Gaonyalelwe Maribe[1]

[1]*Department of Statistics, Faculty of Natural and Agricultural Science, University of Pretoria*

[2]*Department of Statistics, Faculty of Mathematical Science, Ferdowsi University of Mashdad, Iran*

Corresponding author: u15176658@tuks.co.za

data that exceed the capacity of standard analytic tools in terms of volume, velocity, and variety. Although such a wealth of information enables innovation in many disciplines of science, it challenges current statistical and computational methodology, data storage and computational efficiency. We focus on obtaining standard statistical models in scenarios where the capacity of a single computer is surpassed due to the high volume of data. We are particularly interested in linear regression models that address the issues that arise due to multicollinearity. Shrinkage methods are frequently utilized to address the adverse effects of multicollinearity in regression models. Although these methods can easily be applied to small or moderate datasets, they face considerable difficulties in the big data domain. Two of these difficulties are: (a) the size of the data is too large to be loaded into the memory of a computer, and (b) the computational burden is such that the results will not be available in a reasonable time. We propose methods and algorithms for model estimation and validation of closed-form solutions to multiple ridge-type and Liu-type estimators with a general structure that are able to overcome these barriers. Our approach requires minimal access to the entire dataset as it utilizes an array of sufficient statistics that can be computed and updated at row level. The efficiency of our approach is illustrated through an extensive simulation study as well as a real-world application.

Hierarchical Bayesian Spatial Small Area Model for Binary Data Under Spatial Misalignment

Kindie Fenahun Muchie[1], Anthony Kibira Wanjoya[2] and Samuel Musili[2]

[1]Pan African University Institute for Basic Sciences, Technology and Innovation, Nairobi , Kenya

[2]Department of Statistics and Actuarial Sciences, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya

Corresponding author: mkindief@gmail.com

Small area model has become a popular method for producing reliable estimates for small areas. Small area modeling may be carried out via model assisted approaches within the design-based paradigm or model-based approaches. A model assisted design-based inference may be reliable in situations when there are large or medium samples in areas, while if data are sparse, model-based approach may be a necessity. Model based Bayesian analysis methods are becoming popular for their ability to combine information from several sources as well as taking account of uncertainties in the analysis and spatial prediction of spatial data. However, things become more complex when the geographic boundaries of interest are misaligned. Some authors have addressed the problem of misalignment under hierarchical Bayesian approach. In this study, we developed and assessed the performance of non-trivial extension of existing hierarchical Bayesian model for binary data under spatial misalignment. In this study, we developed a spatial hierarchical Bayesian small area model for a binary response variable under spatial misalignment. The developed model is a fusion model, considering both areal level and unit level latent processes. The process models generated from the predictors were used to construct the basis so as to alleviate the well-known problem of collinearity between the true predictor variables and the spatial random process. A simulation study demonstrated that the model has good performance.

Robust Adaptive LASSO and Adaptive E-NET Variable Selection and Regularization in Quantile Regression in the Presence of Collinearity Influential Points

Innocent Mudhombo[1] and Edmore Ranganai[2]

[1] *Vaal University of Technology, Vanderbijlpark, South Africa*

[2] *University of South Africa, Florida Park, South*

Corresponding author: innocentm@vut.ac.za

Collinearity influential observations greatly influence the variable selection and parameter estimation in regression analysis. In this presentation, we propose modifications of adaptive LASSO and adaptive E-NET penalized quantile regression (QR) variable selection procedures to deal with collinearity influential points in variable selection regularization. Although the penalization problem for variable selection has been dealt with extensively in the literature for the QR scenario, many existing variable selection procedures fail to deal with both variable selection and regularization due to adverse effects of collinearity influential points. Our suggested procedures deal with shortcomings of existing variable selection and regularization methods in QR in the presence of collinearity influential observations. The adaptive weights in our proposed procedures are based on the robust weighted quantile regression RIDGE (WQR-RIDGE) estimator. The proposed procedures satisfy oracle properties under regularity conditions. The simulated data show that our suggested QR adaptive variable selection procedures deal with collinearity influential observations better, more so in the robust weighted scenarios than other variable selection procedures.

Continuous time parametric multistate transition models with an application

Henrt Mwambi

University of KwaZulu-Natal, Durban, South Africa

Corresponding author: mwambih@ukzn.ac.za

State transition models are an important methodological development in Statistics. The application areas of such models is vast such as in health, ecology, agriculture, environment to mention a few. In this talk a multistate transition model with application to HIV/AIDS disease progression based on CD4 cell count derived stages is discussed. The model structure is such that it allows for inclusion of the effect of covariate in specific state transitions in the process. The model is also used to model viral load suppression and rebound in a multistate model structure. In addition disease stage sojourn times and survival can be estimated hence allowing interventions to mitigate against transitions to worse stages of the disease. An application will be demonstrated using data from a study in KwaZulu-Natal for individuals infected with HIV allowing for a host of covariates both clinical and non-clinical in nature in addition to individual specific characteristics. In conclusion we found out that multistate transition models are an important tool that can be used to manage chronic diseases such as HIV/AIDS and Cancer both at an individual and population level. They are useful to current approaches of care such as personalized care.

Automated quantification of hydraulic failure in plants using deep learning

Tristan Naidoo[1], Stefan Britz[1] and Glenn Moncrieff[2]

[1] *University Of Cape Town, Cape Town, South Africa*

[2] *South African Environmental Observation Network (SAEON), Cape Town, South Africa*

Corresponding author: mylesnaidoo95@gmail.com

Droughts, exacerbated by anthropogenic climate change, threaten plants through hydraulic failure. This hydraulic failure is caused by the formation of embolisms which block water flow in a plant's xylem conduits. By tracking these failures over time, vulnerability curves (VCs) can be created. These curves hold physiological value and can characterise how vulnerable a plant is to hydraulic failure. However, the creation of these curves is laborious and time consuming. Automating the creation of VCs will allow for the vulnerability of a greater number of plants to be characterised. A standout candidate for automation is the optical vulnerability (OV) method of determining hydraulic failure. To automate this method, embolisms need to be segmented across a sequence of images. This presentation will discuss the automation of the OV method. It will consider three fully convolutional models for the segmentation task, namely U-Net, U-Net with a ResNet34 backbone, and W-Net – a repeated U-Net variant. The dataset used consists of three unique species and four unique leaves, where each leaf has its own sequence of images. Using these leaves, three experiments will be discussed: 1) Can a model generalise across samples from the same leaf? 2) Can a model generalise across different leaves of the same species? 3)Can a model generalise across leaves from different species? The results will be assessed on two levels, firstly, how well a model performs on the segmentation task, and secondly how well VCs are reconstructed from model predictions.

Wind direction prediction of South African windfarms via circular modeling

Najmeh Nakhaeirad[1,2], Andriette Bekker[1] and Mohammad Arashi[1,3]

[1] *University of Pretoria*

[2] *DSI-NRF Centre of Excellence in Mathematical and Statistical Sciences (CoE-MaSS)*

[3] *University of Pretoria, Ferdowsi University of Mashhad*

Corresponding author: najmeh.nakhaeirad@up.ac.za

Wind energy production depends not only on wind speed but also on wind direction. Thus, predicting and estimating the wind direction for sites accurately will enhance measuring the wind energy potential. One of the major challenges is the uncertain nature of wind direction which can be presented through probability distributions. Bayesian analysis can improve the modeling of the wind direction using the contribution of the prior knowledge to update the empirical shreds of evidence. This must align with the nature of the empirical evidence as to whether the data are skew or multimodal or not. So far mixtures of von Mises within the directional statistics domain, are used for modeling wind direction to capture the multimodality nature present in the data. In this paper, due to the skewed and multimodal patterns of wind direction on different sites of the locations understudy, a mixture of multimodal skewed von Mises is proposed for wind direction. Furthermore, a Bayesian analysis is presented to take into account the uncertainty inherent in the proposed wind direction model. A simulation study is conducted to evaluate the performance of the proposed Bayesian model. This proposed model is fitted to datasets of wind direction of Marion Island and two wind farms in South Africa and show the superiority of the approach. The posterior predictive distribution is applied to forecast the wind direction on a wind farm. It is concluded that the proposed model offers an accurate prediction by means of credible intervals.

A new fixed point characterisation based test for the Pareto distribution in the presence of random censoring

Lethani Ndwandwe, James Allison and Jaco Visagie

North-west University

Corresponding author: lethani.ndwandwe@nwu.ac.za

We propose a new goodness-of-fit test for the Pareto type I lifetime distribution in the presence of random right censoring. The test is based on a fixed point characterisation, which is a generalisation of the well known Stein method for the approximation of distributions. The finite sample performance of the new test is evaluated and compared to the modified Cramér von Mises and Kolmogorov-Smirnov tests for different censoring proportions and a variety of alternative lifetime distributions by means of a limited Monte Carlo study. It is found that the new test is competitive compared against the two traditional tests for the majority of alternatives considered.

Extending GPAbin to visualise missing multivariate continuous data

Johané Nienkemper-Swanepoel, Sugnet Lubbe and Niël le Roux

Stellenbosch University, Stellenbosch, South Africa

Corresponding author: nienkemperj@sun.ac.za

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

Some Statistical Challenges in the Analysis of Single-Cell RNASeq Data

Bernard Omolo

Division of Mathematics & Computer Science, University of South Carolina - Upstate, USA

Corresponding author: omolob@ukzn.ac.za

In this study, we review some of the statistical challenges that have been encountered in recent analyses of scRNASeq data. We propose an approach for controlling the Type-I error rate when conducting tests on imputed scRNASeq data. For illustration, we apply the proposed approach to colorectal cancer data from a publicly available database.

Changepoint in randomly spaced time series

Michal Pesta

Charles University, Prague, Czechia

Corresponding author: michal.pesta@mff.cuni.cz

Linear relations, containing measurement errors in input and output data, are considered. Parameters of these so-called errors-in-variables models can change at some unknown moment. The aim is to test whether such an unknown change has occurred or not. For instance, detecting a change in trend for a randomly spaced time series is a special case of the investigated framework. The designed change point tests are shown to be consistent and involve neither nuisance parameters nor tuning constants, which makes the testing procedures effortlessly applicable. A change point estimator is also introduced and its consistency is proved. As a theoretical basis for the developed methods, a weak invariance principle for the smallest singular value of the data matrix is provided, assuming weakly dependent and non-stationary errors. The results are presented in a simulation study, which demonstrates computational efficiency of the techniques. The completely data-driven tests are illustrated through a problem coming insurance.

Climate change detection and attribution: A Bayesian hierarchical approach

Jason Pillay

University Of Pretoria, Pretoria, South Africa

Corresponding author: u18067434@tuks.co.za

While climate change has various ways of presenting itself, variables of interest typically take on the form of temperature change. However, the predictor variables used to detect and attribute climate change include time and/or space, but do not use the available knowledge on climate conditions under a certain forcing scenario. In this paper, we show how said knowledge can provide a more practical view to attribution and detection of climate change on a global scale. We assume a linear regression model of temperature change as a dependent variable and forcings-dependent temperature change as our predictors. We quantify uncertainty in model parameter estimations using a Bayesian approach, discuss assumptions of chosen distributions and incorporate Bayesian inference to obtain a posterior distribution of our model's parameters. We then implement our methodology on acquired global air temperature data and on a controlled sample to verify the method. We will discuss the results of the methodology and evaluate our results and highlight limitations and points for further exploration.

Challenges for using Administrative Data for the compilation of Financial Statistics

Sagaren Pillay

Statistics South Africa, Pretoria, South Africa

Corresponding author: sagarenp@statssa.gov.za

There is a growing trend in developed countries to use administrative data to produce official statistics. The demand for timeous and immediately available data by users is also starting to grow in emerging economies. This demand creates an opportunity to save costs in statistical production and reduce response burden but presents numerous technical, capacity, and methodological challenges for Statistics South Africa (Stats SA). The use of administrative data for producing statistics has numerous advantages over sample survey data. The relative collection costs, lower burden placed on respondents, and better coverage make the usage of administrative data very viable. Statistical agencies can, in many instances, obtain administrative data from various sources at virtually no cost. Further, the risks associated with the usage of data obtained from administrative sources are often minimal or manageable. In this study comparisons are made between data from the Annual Financial Statistics (AFS) Survey and two administrative sources. The first part deals with an analysis of the time series on turnover from the AFS survey and VAT turnover data from the South African Revenue Service (SARS). The second part is a multiple case study of data from businesses in the AFS survey linked to data from businesses in Companies and intellectual Property Commission (CIPC) database.

Modelling representative population mobility for COVID-19 spatial transmission in South Africa

Arminn Potgieter[1], Inger Fabris-rotelli[1], Zaid Kimmie[2], Nontembeko Dudeni-Tlhone[3], Jenny Holloway[3], Charl Janse vsn Rensburg[4], Renate Thiede[1], Pravesh Debba[3], Raeesa Manjoo-Docrat[5], Nada Abdelatif[4] and Sibusiswe Makhanya[6]

[1] *University Of Pretoria*

[2] *Foundation for Human Rights*

[3] *Council for Scientific and Industrial Research*

[4] *Medical Research Council*

[5] *University of the Witwatersrand*

[6] *IBM Research*

Corresponding author: inger.fabris-rotelli@up.ac.za

The COVID-19 pandemic starting in the first half of 2020 has changed the lives of everyone across the world. Reduced mobility was essential due to it being the largest impact possible against the spread of the little understood SARS-CoV-2 virus. To understand the spread, a comprehension of human mobility patterns is needed. The use of mobility data in modelling is thus essential to capture the intrinsic spread through the population. It is necessary to determine to what extent mobility data sources convey the same message of mobility within a region. This paper compares different mobility data sources by constructing spatial weight matrices at a variety of spatial resolutions and further compares the results through hierarchical clustering. We consider four methods for determining connectivity matrices representing mobility between spatial units, taking into account distance between spatial units as well as spatial covariates. This provides insight for the user into which data provides what type of information and in what situations a particular data source is most useful.

Penalized feature selection in model-based clustering

Luandrie Potgieter

University Of Pretoria, Pretoria, South Africa

Corresponding author: luan3potgieter@gmail.com

Cluster analysis is a popular unsupervised statistical method used to group observations into clusters. Clustering helps to identify latent patterns and groupings in data which aids in the understanding of natural phenomena. The data-driven society we live in today has made high dimensional data quite ubiquitous and hence noise variables are unavoidable. Making use of all available variables when modeling can lead to over-parameterization. In addition,high-dimensional data opens the door for the curse of dimensionality. Thus, performing variable selection will ameliorate the model's fit and ease the interpretation of results obtained through clustering. In this presentation, we perform variable selection through penalized model-based clustering. Specifically, an appropriate penalty is chosen to penalize the log-likelihood and the EM algorithm is used to maximize it.

Monitoring procedures for strict stationarity based on the multivariate characteristic function

Charl Pretorius[1], Sangyeol Lee[2] and Simos Meintanis[3]

[1] University Of The Free State, Bloemfontein, South Africa,

[2] Seoul National University, Seoul, South Korea

[3] National and Kapodistrian University of Athens, Athens, Greece

Corresponding author: pretoriusc4@ufs.ac.za

We propose model-free monitoring procedures for strict stationarity of a given data generating process. The new criteria are formulated as L2-type statistics incorporating the multivariate empirical characteristic function. The monitoring procedures are shown to be consistent against nonstationary alternatives, and the null distributions of the monitoring statistics are derived under general conditions which allow for many popular time series models, including stationary ARMA and GARCH models. Results from a numerical study are presented which show that the newly proposed procedures have favourable finite-sample performance when compared to existing monitoring procedures. The talk is concluded with an application in which we test for possible stationarity breaks in financial time series data.

Computational considerations for asymmetric angular- and real data as direction and distance in the modelling of animal movement

Gopika Ramkilawon, Johan Ferreira and Najmeh Nakhaeirad

University Of Pretoria, Pretoria, South Africa

Corresponding author: gopi.ramkilawon@gmail.com

Animal movement is a fundamental part of ecology, and aids in understanding and modelling of social responsibility phenomena including population and community structure dynamics. Movement of animals is often characterised by direction (measured on the circle) and distance (measured on the real line), but traditionally employed models do not account for potential asymmetric angular movement. This study focuses on the modelling of circular data in this animal movement setting using previously unconsidered circular distributions which may allow for a departure from symmetry. In addition, mixtures of often considered models for distance is considered and computational aspects of this joint modelling highlighted. A general hidden state Markov model is used to incorporate both these essential components when estimating via the EM algorithm, and goodness-of-fit measures verifies the validity and viable future consideration of these newly proposed theoretical models within this practical and computational animal movement environment.

LASSO and E-NET Variable Selection and Regularization in Quantile Regression via Minimum Covariance Determinant based Weights

Edmore Ranganai[1] and Innocent Mudhombo[2]

[1] *Unisa, Roodepoort, Florida Park, South Africa*

[2] *Vaal University of Technology, Vanderbijlpark, South Africa*

Corresponding author: rangae@unisa.ac.za

The importance of variable selection and regularization procedures in multiple regression analysis cannot be overemphasized. These procedures are adversely affected by predictor space data aberrations as well as outliers in the response space. To counter the latter, robust statistical procedures such as quantile regression which generalizes the well-known least absolute deviation procedure to all quantile levels have been proposed in the literature. Quantile regression is robust to response variable outliers but very susceptible to outliers in the predictor space (high leverage points) which may alter the eigen-structure of the predictor matrix. High leverage points that alter the eigen-structure of the predictor matrix by creating or hiding collinearity are referred to as collinearity in?uential points. In this paper, we suggest generalizing the penalized weighted least absolute deviation to all quantile levels, i.e., to penalized weighted quantile regression using the RIDGE, LASSO, and elastic net penalties as a remedy against collinearity in?uential points and high leverage points in general. To maintain robustness, we make use of very robust weights based on the computationally intensive high breakdown minimum covariance determinant. Simulations and applications to well-known data sets from the literature show an improvement in variable selection and regularization due to the robust weighting formulation.

Shared component modelling of early childhood anaemia and malaria in four sub-Saharan African countries

Danielle Roberts and Temesgen Zewotir

University of KwaZulu-Natal, Durban, South Africa

Corresponding author: robertsd@ukzn.ac.za

Malaria and anaemia contribute substantially to child morbidity and mortality. Using a child-level shared component model, we sought to jointly model the residual spatial variation in the likelihood of these two correlated diseases, while controlling for individual-level, household-level and environmental characteristics. This shared component model allowed the district-level spatial effect to be partitioned into a shared and disease-specific spatial component. The results indicated that the spatial variation in the likelihood of malaria was more prominent compared to that of anaemia, for both the shared and specific spatial components. In addition, multiple districts associated with an increased likelihood of anaemia but a decreased likelihood of malaria were identified. This suggests that there are other drivers of anaemia in children in these districts, which warrants further investigation. The maps of the shared and disease-specific spatial patterns provide a tool to allow for more targeted action in malaria and anaemia control and prevention, as well as for the targeted allocation of limited district health system resources.

Identifying rare cell types among large and diverse populations of immune cells with precision and robustness

Miguel Rodo[1,2], Virginie Rozot[2], Carly Young[2], Phu Van[3], Evan Greene[3], Greg Finak[3], Raphael Gottardo[3], Francesca Little[1] and Thomas Scriba[2]

[1] *Department of Statistical Sciences, University of Cape Town, Cape Town, South Africa*

[2] *South African Tuberculosis Vaccine Initiative, Institute of Infectious Disease and Molecular Medicine, Division of Immunology, Department of Pathology, University of Cape Town, Cape Town, South Africa*

[3] *Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, United States of America*

Corresponding author: rdxmig002@myuct.ac.za

High-dimensional analysis of immune responses informs drug and vaccine design for combating diseases, such as tuberculosis and COVID-19. We consider the problem of identifying and characterising small cell subsets within a diverse population of immune cells based on increased protein production in response to a pathogen. Classifying cells as responding is a difficult problem, as downstream analyses depend on sensitive detection for statistical power, but classification needs to be highly specific as responding cells are typically very infrequent. This can be compounded by measurement batch effects between experiments, and protein distributions overlapping between responding and non-responding cells. Traditionally, classification has been performed by manually drawing lines, but this is time-consuming, computationally irreproducible and prone to subjectivity and lapses in judgment. Previous automated methods use uninterpretable tuning parameters requiring precise settings, do not allow for overlap between responding and non-responding cells and/or assume no batch effects. We propose a simple Empirical Bayes approach, based on the two-groups model. It achieves strong concordance with manually clustered samples, in terms of cell classification and downstream clinical results, in significantly less time. Simulation shows that it is robust to batch effects, performs equivalently across a range of interpretable tuning parameters, and overcomes overlap in the measurement distribution between responding and non-responding cells. Finally, we apply it to a new high-dimensional immunological dataset and discover a novel cell subset associated with tuberculosis disease development. In conclusion, this approach is faster than manual classification, and out-performs existing automated methods.

An Algorithm for Generating Multi-Label Classification Data

Trudie Sandrock

Stellenbosch University

Corresponding author: trudies@sun.ac.za

Multi-label classification has become an active area of research. When comparing multi-label classification methods, benchmark datasets are generally used. However, these benchmark datasets have significant shortcomings and ideally methods should be compared using artificially generated data, however, to date, few proposals exist in this regard and the existing proposals are limited in many regards. A new method for generating multi-label classification data is therefore proposed, which offers considerable control over many properties of the simulated data. Of special interest is the option of specifying locally and globally relevant input variables.

Marginalized Two-part Joint Models for Generalized Gamma Family of Distributions

Mohadeseh Shojaei and Ding-Geng (Din) Chen

Department of Statistics, University of Pretoria, Pretoria, South Africa

Corresponding author: m_shojai82@yahoo.com

Positive continuous outcomes with a substantial number of zero values and incomplete longitudinal follow-up are quite common in medical cost data. To jointly model emi-continuous longitudinal data and survival data and to provide marginalized covariate effect estimates, marginalized two-part joint model (MTJM) have been developed for outcomes with lognormal distributions. In this paper, we propose MTJM models for outcomes from a generalized gamma (GG) family of distributions. The GG distribution constitute an extensive family that contains nearly all of the most commonly used distributions including the gamma, exponential, Weibull and log normal. In the proposed MTJM-GG model, the conditional mean from a two-part model with a three-parameter GG distribution is parameterized to provide that marginal interpretation for regression coefficients. MTJM-gamma and MTJM-Weibull are developed as special cases of MTJM-GG. To illustrate the applicability of the MTJM-GG, we applied the model to a set of real electronic health record data recently collected in Iran and we provide SAS code for implementation. The simulation results show that when the response distribution is unknown or mis-specified, which is usually the case in real data sets, the MTP-GG is preferable to other models. The advantage of using the GG family of distribution is that it facilitates estimating a model with improved fit over the standard Weibull or log-normal distributions.

A new double sampling scheme to monitor the process mean of autocorrelated observations using an AR(1) model with a skip sampling strategy

Sandile Shongwe

University of the Free State

Corresponding author: ShongweSC@ufs.ac.za

There are a lot of academic research on statistical process monitoring schemes that assume that sequential observations are independent and identically distributed (iid), however, in industrial processes, sequential data tends to exhibit serial correlation (i.e. autocorrelation). Implementing monitoring schemes designed for iid observations when in fact data is sampled from an autocorrelated process yields misleading results. In this paper, we propose a side-sensitive double sampling (SSDS) scheme to monitor the mean of autocorrelated observations using a first-order autoregressive model. In order to reduce the negative effect of serial dependence, a sampling strategy that involves sampling of non-neighbouring observations (i.e., skipping s observations before sampling) is incorporated into the computation of the probability values of the run-length distribution and the charting limits. The main findings of this study is that the proposed s-skip SSDS scheme yields a run-length distribution that has uniformly better average run-length (ARL) and expected ARL values as compared to the existing non-side-sensitive double sampling scheme and other well established Shewhart-type schemes (i.e. runs-rules and synthetic) for autocorrelated observations. A real life example for yogurt cup filling process is used to illustrate how the proposed monitoring scheme is implemented.

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

Isaac Singini

University of Pretoria, Hatfield, Pretoria, South Africa

Corresponding author: isaac.singini@up.ac.za

Joint models for longitudinal and survival data are a class of models that jointly analyse an outcome repeatedly observed over time such as a bio-marker and associated event times. There are two main classes of these models namely, shared parameter and latent class joint models. The main difference between these two modeling frameworks is that latent class joint models make no assumption about the association between the time-dependent covariate(s) and risk for an event while shared parameter joint models do not explicitly handle heterogeneity in the population These models are useful in two practical applications, firstly focusing on survival outcome whilst accounting for time varying covariates measured with error and secondly focusing on the longitudinal outcome while controlling for informative censoring. Interest on the estimation of these joint models has grown in the past two and half decades with minimal effort directed towards developing influence diagnostics. In this study we compared Cook's statistics for detecting influential subjects to classes identified by the latent class joint model which in effect would classify influential subjects through population heterogeneity. This approach was illustrated using data from a multi-center clinical trial on TB pericarditis. The data confirmed our hypothesis that latent class joint models can be used to as an alternative to diagnostics to identify influential subjects in the shared parameter joint models for longitudinal and survival data. This is done by classifying heterogeneous classes using a latent variable.

Modelling Aleatory and Epistemic Uncertainty in Natural Hazard Distributions

Ansie Smit[1] and Alfred Stein[2]

[1] *University Of Pretoria Natural Hazard Centre, Pretoria, South Africa*

[2] *Faculty of Geo-information Science & Earth Observation (ITC), University of Twente, Enschede, Netherlands*

Corresponding author: ansie.smit@up.ac.za

A versatile method to assess natural hazards that accounts for both epistemic and aleatory uncertainty in natural hazard distributions is presented. The events in observed datasets of natural phenomena can be classified as prehistoric, historic or instrumentally recorded data. Each of these types of datasets exhibits different types of epistemic and/or aleatory uncertainties. The described methodology accounts for incomplete datasets, uncertainty associated with the observed event sizes, the applied distributions, and with the validity of occurrence of events in the dataset. These types of uncertainty are addressed using convolution and mixture distributions, and weighted likelihood functions. The different data types are combined using likelihood functions which allows for the maximum likelihood (ML) estimation and Bayesian inference (BI) of the parameters. The methodology is tested on a synthetic natural hazard dataset, with various combinations of uncertainty investigated. Estimates of the parameters yielded markedly different results, with BI providing overall more precise estimates than MLE. This in turn can have a large effect on estimates of the return periods of event sizes of natural hazards.

The Practical Considerations of a Flexible Finite Mixture Regression (FMRFLEX) Framework.

Riaan Smit[1,2], Sollie Millard[2] and Frans Kanfer[2]

[1] *Vodacom, Noordwyk, South Africa*

[2] *University of Pretoria, Hatfield, South Africa*

Corresponding author: u21150177@tuks.co.za

Finite mixture regression (FMR) models represent a flexible statistical modelling framework which allows for the underlying structure of complex heterogeneous datasets to be quantified and in doing so, offer increased predictive power compared to traditional one-class regression models. In practice, a heavy reliance is placed on linear models as the common input predictors in finite mixture regression models. To improve on the overall predictability of these models, Ahonen et al (2019) proposed a revised formulation of the finite mixture regression methodology (FMRFLEX) which introduces a more flexible structure to the linear predictors included in these models. This flexibility in the structure of the linear predictors is achieved through the combination of a random forest learner and lasso-penalized finite mixture regression model. By following this approach, the nonlinearities and interactions inherent in the data is “captured” by the random forest learner in a flexible and data-driven manner, which is in turn combined with the linear associations derived from the original covariates using standard penalized finite mixture regression methods. Empirical results have shown that the FMRFLEX model proposed by Ahonen et al (2019) is able to achieve greater predictability compared to traditional finite mixture regression models when some of the regression components inherent in a specific dataset, is nonlinear. The practical implementation of the FMRFLEX methodology will be presented with special attention given to the predictive performance of the model.

Empowering differential networks using Bayesian analysis

Jarod Smith[1], Mohammad Arashi[1,2] and Andriëtte Bekker[1]

[1] *Department of Statistics, University Of Pretoria, Pretoria, South Africa*

[2] *Department of Statistics, University Of Pretoria, Pretoria, South Africa & Department of Statistics, Faculty of Mathematical Sciences, Ferdowsi University of Mashhad, Mashhad, Iran*

Corresponding author: jarodsmith706@gmail.com

Differential networks (DN) are important tools for modeling the changes in conditional dependencies between multiple samples. A Bayesian approach for estimating DNs, from the classical viewpoint, is introduced with a computationally efficient threshold selection for graphical model determination. The algorithm separately estimates the precision matrices of the DN using the Bayesian adaptive graphical lasso procedure. Synthetic experiments illustrate that the Bayesian DN performs exceptionally well in numerical accuracy and graphical structure determination in comparison to state of the art methods. The proposed method is applied to South African COVID-19 data to investigate the change in DN structure between various phases of the pandemic.

An edge preserving median filter for images based on level-sets

JP Stander, Inger Fabris-rotelli, Theodore Loots, JM van Niekerk and Alfred Stein

University Of Pretoria

Corresponding author: inger.fabris-rotelli@up.ac.za

In this article, we propose an edge-preserving median filter for noise removal in images. This filter uses connected sets of pixels of the same value to determine flexible regions which contour to edges in the image. The filter determines whether a set is noise or signal and smooths this noise. These regions are flexible since they are created based on their values, namely are data-driven and therefore provide the mechanism for the filter to preserve edges in the image. Current median filters do not preserve edges. Using metrics such as Pratt's Figure of Merit and Peak Signal to Noise Ratio on example images from the labeled faces in the wild data set was concluded that the proposed filter does remove noise while preserving the edges in the image.

Multiscale decomposition of spatial lattice data for feature detection

Rene Stander, Inger Fabris-rotelli, Ding Chen and Gregory Breetzke

University Of Pretoria

Corresponding author: inger.fabris-rotelli@up.ac.za

The Discrete Pulse Transform (DPT) has only been applied to signals in 1D and 2D on regular lattices. The theory leaves scope for the application on irregular spatial lattice data in 2D, also referred to as areal data in spatial literature. In this paper, we extend the DPT theory for irregular lattice data as well as consider its efficient implementation, the Roadmaker's Pavage, and visualisation. The DPT was derived considering all possible connectivities satisfying the morphological definition of connection. Our implementation allows for any connectivity applicable for regular and irregular lattices. We present the implementation of the Roadmaker's Pavage algorithm on spatial images as well as irregular lattice data with a toy example and illustrate this with two applications. The theory is applied to brain imagery (regular lattice) as well as crime counts (irregular lattice data) for feature detection. Using the multiscale Ht-index as a measure of saliency on the extracted DPT pulses, important features from both the regular and irregular lattice data can be detected.

Shrinkage methods for the estimation of the extreme value index

Matthys Lucas Steyn[1,2], Tertius de Wet[1], Bernard De Baets[2] and Stijn Luca[2]

[1] *Stellenbosch University, Stellenbosch, South Africa*

[2] *Ghent University, Ghent, Belgium*

Corresponding author: lucasteyn@sun.ac.za

A fundamental problem in extreme value analysis is the estimation of the extreme value index (EVI), denoted by γ. The EVI characterises the rate at which the tails of a distribution decay which, in turn, enables the estimation of extreme quantiles or excess probabilities. A common approach to estimate the EVI is the Hill estimator for the $\gamma > 0$ case and the generalised Hill estimator for the case of a real-valued EVI. Under a second-order condition of the extreme value theorem, these estimators have been expanded to a regression model that reduces the bias inherent to the Hill-type estimators. It has previously been proposed to use a ridge penalty to reduce the variance of the estimators under the bias-reduced model. We present shrinkage methods for the estimation of the EVI under the second-order condition. Specifically, estimators of the EVI under the non-linear regression model with the L1 and L2 penalties are presented. The asymptotic properties of the estimators under the penalised regression model are discussed. These methods are compared on simulated and real-world data sets to demonstrate the advantage of using a regularised, second-order approach to estimate the EVI.

Machine Learning techniques illustrated using R-Shiny

Lourens Strydom

University Of Pretoria, Pretoria, South Africa

Corresponding author: u17341745@tuks.co.za

Innovating and exciting models that converge to the pinnacle of human intelligence are at the frontier of research nowadays. This project will provide the reader with fundamental concepts used daily in Machine Learning. From well-established algorithms that have been operating for decades on end to newer and stronger algorithms in full fruition. It is thus fitting to introduce the reader to a modern web application fabricator, R-Shiny, that displays stunning visualizations and accomplishes powerful analysis by using the vigorous abilities of R. Machine Learning is well covered in academics. But textbooks have never done it justice. By allowing students to interact with some of the algorithms, we hope to inspire them. Until recently, R users had no experience in web development. R-Shiny has changed that. It is now possible to dynamically utilize your code for daily tasks with action buttons, sliders, selection lists, and many more features.

A stochastic network model for estimating population mobility between areal units in an irregular lattice

Renate Thiede[1], Inger Fabris-Rotelli[1], Pravesh Debba[2] and Christopher Cleghorn[3]

[1]*Department of Statistics, University Of Pretoria*

[2]*Council for Scientific and Industrial Research Pretoria, Department of Statistics and Actuarial Science, University of the Witwatersrand*

[3]*School of Computer Science and Applied Mathematics, University of the Witwatersrand, Johannesburg, South Africa*

Corresponding author: renate.thiede@gmail.com

Modelling population mobility is essential for many applications, including urban planning, contact tracing and access to facilities. In particular, it is relevant to model how people move between discrete spatial units, such as municipalities or provinces, modelled in spatial statistics as irregular lattices. Complex road networks are well suited to model mobility, however, modelling the road network of a region as a whole is computationally expensive. This paper models the movement of people between spatial units, particularly electoral wards, in a representative, computationally feasible manner. Mobility is modelled as a Markov chain, with the wards as states. To simplify the road network, Louvain clustering is used to select representative nodes as entry points into the network in each ward. One-step transition probabilities are obtained by calculating the probability of moving from one of the representative nodes in a ward into any of the representative points in a spatially adjacent ward. For each ward, we obtain a matrix of probabilities of moving from that ward to any other ward in the study area. Only transitions into spatially adjacent neighbours may have a positive probability, while transitions to non-adjacent wards will have a probability of zero, ensuring that the one-step mobility matrices are sparse. To obtain the probabilities of journeys that cross multiple wards, we multiply the relevant sparse one-step transition matrices. This provides a computationally simple approach to model population mobility, resulting in mobility matrices that can be used as input in spatial epidemiological models, accessibility analyses and other spatial models.

Sales Forecasting Using Linguistic Fuzzy Logic with Weather Data

Tomáš Tichý

VSB-TUO, Ostrava, Czech Republic

Corresponding author: tomas.tichy@vsb.cz

This text proposes a novel approach studying financia quantities using various exogenous variables and is inspired by fuzzy natural logic. The method is based on modeling the influence of exogenous variables on financial quantities by fuzzy linguistic IF-THEN rules. Reliable estimation of customer demand for products and services constitutes a key aspect of financial planning in every company. For example, when estimating future sales, as a proxy to demand, in addition to pure economic quantities, a large selection of (exogenous) variables specific to a given product can be considered. Potential impact of weather conditions on sales has been known for very long time, though the research using weather data has been mostly focused on energy sector. As concerns retail, several authors have started to analyze this issue only recently. The proposed methodology is applied to real sales data and compared with a standard approach. The results are promising especially when frequently collected weather data are considered, even if sales are collected in longer periods.

A Quantitative Analysis of Investor Over-reaction and Under-reaction in the South African Equity Market: A Fuzzy C-Mean Algorithm

Aude Ines Tiekwe, Willie Conradie and Rousseau Lötter

Stellenbosch University, Stellenbosch

Corresponding author: ines@aims.edu.gh

One of the basic foundations of traditional finance is the theory underlying the efficient market hypothesis (EMH). The EMH states that stocks are fairly and accurately priced, making it impossible for investors to use stock selection, technical analysis, or market timing to outperform the market by earning abnormal returns. Several schools of thought have challenged the EMH by presenting empirical evidence of market anomalies, which seems to contradict the EMH. One such school of thought is behavioural finance, which holds that investors over-react and/or under-react over time, driven by their behavioural biases. In this study, a Fuzzy c-Means Model, based on the technique of pattern recognition is used to investigate investor's over-reaction and under-reaction in the South African equity market. The study used quarterly data of 163 shares in the Johannesburg Stock Exchange AllShare index, selected from the top 100 shares listed for the period 2006 to 2016 and downloaded from Iress and Bloomberg. Over-reaction and under-reaction were both detected, and differed across sectors. No clear patterns of the two biases investigated were visible over time. The results of the FCM analysis revealed that the resources sector shows the most under-reaction. The results of this study imply that momentum and a contrarian investment strategies can lead to over-performance in the South African equity market, but can also generate under-performance in a poorly performing market. Therefore, no trading strategies can be advised based on the results of this study.

Evaluating the Effect of HIV Status Awareness on HIV Risky Sexual Behaviours and Marriage dissolution using Marginal Structural Models

Halima Twabi[1], Samuel Manda[2], Dylan Small[2] and Hans-Peter Kohler[2]

[1] *Department of Mathematical sciences, University of Malawi, Zomba, Malawi*

[2] *South African Medical Research Council*

Corresponding author: htwabi@unima.ac.mw

Knowledge of HIV status has been shown to impact risky sexual behaviors such as inconsistent condom use and multiple sexual partners and on marriage dissolution. Increase in risky sexual behaviors results in a high HIV transmission. Policy makers would be interested to assess the magnitude of HIV status awareness and its impact on risky sexual behavior and marriage outcomes. This paper aimed to use routine longitudinal data to estimate the effect of HIV status awareness on risky sexual behaviors and marriage dissolution. Data was extracted from the Malawi Longitudinal Study for Families and Health (MLSFH) and complete linked individual data that appeared in 8 waves collected bi-annually was used. A Marginal Structural Model (MSM) using the inverse probability of treatment weights (IPTW) was used to estimate the known HIV-status effect on consistent condom use, multiple sexual partners and marriage outcomes. The findings of the study show that HIV-status awareness had a beneficial effect on condom use and having multiple sexual partners. However, there was an increase in marriage dissolution among individuals who were aware of their HIV-positive status. The study may suggest effectiveness of HIV preventive strategies in Malawi. We recommend continuation of interventions that promote HIV testing and counselling to help people become aware of their HIV status.

Forward stagewise linear regression for ensemble methods

Danie Uys

Stellenbosch University, Stellenbosch, South Africa

Corresponding author: dwu@sun.ac.za

In supervised learning, the forward stagewise regression algorithm is considered a more constrained version of forward stepwise regression. In its turn, the forward stagewise regression algorithm can be refined to produce the incremental forward stagewise regression model. In the latter model, the idea of slow learning is introduce where the residual vector and the appropriate regression coefficient are updated in very small steps at each iteration. Ensemble methods combine a large number of simpler base learners to form a collective model that can be used for prediction. Learning methods such as Bagging, Random Forests and Boosting amongst others, can all be regarded as ensemble methods. In these methods, the linear model is expressed as a linear combination of these simpler base learners, where the coefficients of the base learners are to be estimated by least squares. Since a large number of base learners is typically involved, the residual sum of squares of the linear combination of base learners has to be penalised by, for example, the lasso penalty. However, the large number of base learners also complicates the minimisation of the coefficients in the penalised residual sum of squares criterion. By using the iterative forward stagewise linear regression algorithm for ensemble methods, which includes the idea of slow learning and closely approximate the lasso, estimators of the coefficients of the base learners can be obtained. In the talk, the performance of various ensemble methods is evaluated. This is done by applying the forward stagewise linear regression algorithm for ensemble methods to simulated, as well as real life datasets.

Identification of Latent Growth Classes in a South African Birth Cohort study

Noëlle Van Biljon[1], Marilyn Lake[2], Lesley Workman[2], Heather Zar[2] and Francesca Little[1]

[1] *Department of Statistical Sciences, University Of Cape Town, Cape Town, South Africa*

[2] *Department of Paediatrics and Child Health, University Of Cape Town, Cape Town, South Africa*

Corresponding author: anbnoe001@myuct.ac.za

Numerous methods are available to model and analyse longitudinal growth data. Conventionally, such growth modelling methods focus on the analysis of average longitudinal trends or identify those belonging to groups of abnormal growth based on standardised z-scores, in addition to investigating potential predictors of abnormal growth. Latent Class Mixed Modelling (LCMM) allows identification of groups of subjects that follow similar longitudinal trends, be they normal or abnormal, based on a combination of a linear mixed-effect, structural equation and multinomial logistic modelling. Here LCMM was used to identify underlying latent profiles of growth for height, weight, head circumference (HC), mid-upper arm circumference (MUAC), triceps skin fold thickness (TRI), body mass index (BMI) and weight for height (WFH) measurements taken from birth until the age of five years for a sample of 1143 children from the Drakenstein Child Health Study (DCHS). Subsequently, three classes of growth within height (n_1=42, n_2=664, n_3=425), weight (n_1=606, n_2=455, n_3=72), HC (n_1=684, n_2=404, n_3=42), MUAC (n_1=58, n_2=241, n_3=710), BMI (n_1=673, n_2=185, n_3=273) and WFH (n_1=203, n_2=778, n_3=93), each with distinct trajectories over childhood were identified and validated. With the identification of these classes, a better understanding of distinct childhood growth trajectories and their predictors may be distinguished, informing interventions to promote optimal childhood growth.

Robust mixture regression with unspecified error distributions

Wihan van der Heever, Sollie Millard and Frans Kanfer

University of Pretoria, Pretoria, South Africa

Corresponding author: wihanvdheever@gmail.com

Mixture of linear regression models have become innate when fitting a response variable to one or more feature variables in the face of latent components present in data. These models frequently employ maximum likelihood estimation (MLE) for the regression parameters and depend on the assumption that the errors of the underlying components are normally distributed. Naturally, when this assumption is breached, the traditional approach to mixture regression is no longer viable for modelling purposes, due to the bias transpiring from the model not capturing the ad rem relationship between variables. This paper considers a semiparametric mixture of linear regression model, which reduces the bias by introducing a kernel density-based expectation maximisation (KDEEM) algorithm. This algorithm accommodates linear mixture regressions without specifying the component error distributions, thereby allaying the complications arising from the violation of the normality assumption. The paper uses a simulation study to compare the KDEEM approach to standard MLE in cases of normally distributed and non-normally distributed errors. The KDEEM algorithm is also applied to a practical Covid-19 data set.

A noncentral Lindley construction illustrated illustrated in an INAR(1) environment

Ané Van Der Merwe and Johan Ferreira

University Of Pretoria, Department Of Statistics, Hatfield

Corresponding author: ane.neethling@up.ac.za

This study proposes a previously unconsidered generalization of the Lindley distribution by allowing for a measure of noncentraility (ncL). Essentially structural properties are investigated and derived in explicit andtractable forms, and the estimability of the model is illustrated via real data. This distribution is then used as a candidate for the rate parameter of the Poisson distribution, which allows for departure from the usual equidispersion restriction of the Poisson distribution when modelling count data. This Poisson-noncentral Lindley (PncL) is also systematically investigated and characteristics are derived. The impact of this model is illustrated in both a simulation study as well as real data, by implementing this PncL model as the count error model in an integer autoregressive (INAR) model. The effect of this systematically-induced noncentrality parameter is illustrated and paved the way for future flexible modeling, not only as a stand-alone contender in Lindley-type scenarios (as the ncL) but also in discrete time series scenarios (as the PncL) when the often-assumed equidispersed assumption is not adhered to in practical data environments.

Exploding biplots with density axes in Plotly

Carel Johannes Van Der Merwe, Delia Sandilands and Sugnet Gardner-Lubbe

Stellenbosch University, Stellenbosch, South Africa

Corresponding author: cjvdmerwe@sun.ac.za

Biplots are useful when visualizing multivariate data. It can, however, sometimes be challenging to interpret, for example when the axes and points cause overcrowding of the plot. This overcrowding is often due to the presence of many variables, highly correlated variables, or merely data sets with a large number of observations. In this paper improvements to the biplot are made to address these shortcomings. These improvements include: i) the automatic parallel translation, or “explosion”, of axes, ii) the use of densities on the axes to improve interpretation and representation of large data sets, and iii) introducing interactive biplots via the use of the Plotly package in R. These improvements result in a better composition of the plot to make it seem less crowded, more easily interpretable, offer additional information that can get lost in the case of a high volume of data, and allowing the user to inspect the biplot element-wise. An accompanying Shiny web-based application was also created and is available at https://carelvdmerwe.shinyapps.io/ExplodingBiplots/.

Multivariate prediction with machine learning in digital soil mapping

Stephan Van Der Westhuizen[1,2], Gerard BM Heuvelink[2], David Hofmeyr[1] and Laura Poggio[3]

[1] *Stellenbosch University, Stellenbosch, South Africa*

[2] *Wageningen University & Research, Wageningen, The Netherlands*

[3] *ISRIC - World Soil Information, Wageningen, The Netherlands*

Corresponding author: stephanvwd@sun.ac.za

Soil maps produced with digital soil mapping (DSM) contain vital information about the spatial distribution of soil properties which are used in fields such as agronomy and ecology. DSM uses statistical models to quantify the relationship between a soil property and a selection of soil-forming representative environmental covariates, and then uses this relationship to predict the soil property at locations where it was not observed. Soil maps are usually produced with univariate statistical models, i.e. each map is produced independently from others without taking into account the underlying correlation structure between the soil properties. This can lead to inconsistent predictions, for example, mapping soil organic C and N concentrations with separate univariate models may lead to unrealistic C:N ratios. Many examples of multivariate mapping exist, and co-kriging is probably the most widely used multivariate technique in DSM. However, co-kriging requires severe restrictions such as the linear model of coregionalisation. Machine learning applications in DSM has gained tremendous popularity over the last decade, but the use of machine learning to perform multivariate mapping in DSM is still lacking. In this presentation we compare the multivariate extensions of random forests and projection pursuit regression to (regression) co-kriging when predicting two soil properties, organic C and N, from the Land Use and Coverage Area Survey (LUCAS) data set. Maintaining a well represented C:N ratio is important for map users as it provides information on the residue decomposition and the nitrogen cycle in soil.

Competing risks joint models using R-INLA

Janet van Niekerk

King Abdullah University of Science and Technology

Corresponding author: janet.vanniekerk@kaust.edu.sa

In this talk, we introduce a framework based on R-INLA to apply competing risks joint models in a unifying way such that non-Gaussian longitudinal data, spatial structures, times-dependent splines and various latent association structures, to mention a few, are all embraced in our approach. Our motivation stems from the SANAD trial which exhibits non-linear longitudinal trajectories and competing risks for failure of treatment.

The influence of different regimes on the estimation of GARCH volatility parameter estimation

Lienki Viljoen, Monique-Mari Britz and Willie Conradie

Stellenbosch University, Stellenbosch, South Africa

Corresponding author: lienki@sun.ac.za

Volatility is used as a measure of risk within the financial markets. GARCH modelling involves important volatility forecasting methodology and is widely used in finance. It is important to be able to forecast volatility since volatility has an impact on financial portfolios and the risk hedging methodology followed by financial companies. The parameter estimates and volatility forecasts of three GARCH models, the Symmetric GARCH, GJR-GARCH and E-GARCH models, are compared using the JSE All-Share index. This index is divided into two different periods, namely, a tranquil financial period and a turbulent financial period. Different factors influence the performance of GARCH models and consequently determines which GARCH model is the most suited for certain circumstances. These factors are: the window period, forecasting horizon, the financial period and the underlying distribution of the log returns.

On an omnibus test for the parametric Cox proportional hazards model

Jaco Visagie, James Allison, Elzanie Bothma and Marius Smuts

North-West University

Corresponding author: jaco.visagie@nwu.ac.za

We propose an omnibus test of fit for the parametric Cox proportional hazards model in the presence of random right censoring. The proposed test results from a modification of an existing test for the uniform distribution. This test is demonstrated to be able to detect deviations from the hypothesised model in two cases, first when the baseline distribution is misspecified and second when the regression component of the model is misspecified. Two modified classical tests are considered and a Monte Carlo study shows that the newly proposed test outperforms these tests for the majority of alternatives included. As a result of independent interest, we outline the procedure required to use the newly modified test in the framework of independent and identically distributed random variables.

Portfolios and interviews: Turning our students into lifelong learners

Michael von Maltitz

University of the Free State, Bloemfontein, South Africa

Corresponding author: vmaltitzmj@ufs.ac.za

Over the past two years I have experimented with a form of teaching and assessment that is considered novel (and possibly risky) in the statistics education field. This system is based on Fink's taxonomy and completely authentic learning, aiming to turn students into curious, lifelong learners. I use realistic assessments to eliminate the text-to-test (or cram-and-forget) mentality, to encourage deeper learning, and to pass on critical life skills. I will introduce the pedagogical evolution that led me to adopt these methods, the practical implementation of the system, the advantages and disadvantages of these methods in statistical education, and the insight gained from experimenting with this process so far.

Robust mixture regression using mean-shift penalisation

Anika Wessels, Frans Kanfer and Sollie Millard

University of Pretoria, Pretoria, South Africa

Corresponding author: anikawessels92@gmail.com

Finite mixture regression models the relationship between a response variable and feature variables in the presence of latent groups in the population. The regression model parameters are unique to each latent group quantifying the different regression structures. Although the classical normal mixture regression model is mostly used since it simplifies the estimation and interpretation, it can be highly sensitive to outliers present in the data. Failing to account for this may distort the results and lead to inappropriate conclusions. We consider a mean-shift robust mixture regression approach. This method uses a sparse, component specific and scale dependent mean-shift parameterisation to simultaneously identify the outliers and perform robust parameter estimation. The properties of the technique are demonstrated using a simulation study.

Semiparametric modelling for diabetic retinopathy among type II Diabetic Patients

Bezalem Eshetu Yirdaw

University Of South Africa, South Africa

Corresponding author: 12962805@mylife.unisa.ac.za

The proportion of patients with diabetic retinopathy has grown with increasing number of diabetic mellitus patients in the world. It is among the top risk factors of blindness worldwide, especially those living in developing countries. The main objective of this study was to identify contributing risk factors of diabetic retinopathy among Type II diabetic patients. A sample of 192 patients was selected using systematic random sampling from Black Lion Specialized Hospital diabetic unit from 1 March 2021 to 1 April 2021. A multivariate stochastic regression imputation technique was applied to impute the missing values. The response variable, Diabetic retinopathy is a categorical variable with two outcomes. Plots from univariate analysis showed that duration of diabetes and haemoglobin A1C have a nonlinear relationship with diabetic retinopathy. Therefore, we proposed a semiparametric model, in particular using spline smoothing to analyze the diabetic retinopathy data efficiently. In the multivariate analysis, the statistical test indicated that the spline effects of duration of diabetes and haemoglobin A1C are significant, but the spline effect of cholesterol level was nonsignificant. The model was refitted considering a linear cholesterol level effect. The results revealed that the clinical variables of a type II diabetic patient have strong predictive factors of diabetic retinopathy. Hence, health care workers should be cautious about the possible effects and complications of diabetic mellitus which can be caused by the clinical variables.

Application of Semiparametric Model in modelling Diabetic Retinopathy Among Type II Diabetic Patients at Black Lion Specialized Hospital Addis Ababa, Ethiopia

Bezalem Eshetu Yirdaw and Legesse Kassa Debusho

University Of South Africa, Johannesburg, South Africa

Corresponding author: 12962805@mylife.unisa.ac.za

The proportion of patients with diabetic retinopathy has grown with an increment of diabetic mellitus in the world. It is among the top risk factors of blindness worldwide, especially those living in developing countries. The main objective of this study was to identify contributing risk factors of diabetic retinopathy among Type II diabetic patients. A sample of 192 patients was selected using systematic random sampling from Black Lion Specialized Hospital diabetic unit from 1 March 2021 to 1 April 2021. A multivariate stochastic regression imputation technique was applied to impute the missing values. The response variable, Diabetic retinopathy is a categorical variable with two outcomes. Plots from univariate analysis showed that duration of diabetes and haemoglobin A1C have nonlinear relationship with diabetic retinopathy. Therefore, we proposed a semiparametric model, in particular using spline smoothing to analyse efficiently the diabetic retinopathy data. In the multivariate analysis, the statistical test indicated that the spline effects of duration of diabetes and haemoglobin A1C are significant, but the spline effect of cholesterol level was nonsignificant. The model was refitted considering a linear cholesterol level effect. The study revealed that gender, hypertension, insulin treatment, duration of diabetes and haemoglobin A1C are major determinant factors of diabetic retinopathy whereas controlling the effects of other variables the cholesterol level of a patient did not have a significant effect on diabetic retinopathy.

Posters

Classification and Clustering-based Methods for Outlier Detection of Solar Resource Data

Waldo Abrahams

Department of Statistics, Nelson Mandela University, Gqeberha, South Africa

Corresponding author: s216038413@mandela.ac.za

Almost 90% of the primary global energy demand serviced from the burning of fossil fuels (Abas, Kalair & Khan, 2015). Owing to the detrimental environmental impact of this, a global energy transition to the use of renewable energy, including solar energy, is needed (Gielen et al., 2019). An important aspect that inhibits the growth of solar energy, is accurate solar resource data. Such data is needed because knowledge of the future reliability and quality of energy production is required to analyse a system's performance and determine financial implications (Sengupta et al., 2017). Current methods used to detect outliers in solar resource data do not efficiently identify outliers and an accurate and robust approach is needed. Using simulated and real-world data, this study investigates the use of several classification methods, along with a two-stage clustering-classification approach to accurately identify outliers in solar resource data.

Bootstrap-based tolerance intervals for nested two-way random effects models

Christopher Erasmus

Nelson Mandela University, Gqeberha, South Africa

Corresponding author: christophererasmus1997@gmail.com

Variance component, or random effects models, are frequently used by manufacturers to model the variance present in a manufacturing process. By applying tolerance intervals to variance component models, manufacturers are able to set upper and lower limits to monitor the variance within a process. Existing methods for constructing tolerance intervals are constrained by the necessity for data to be normally distributed. Recently, non-parametric bootstrap-based techniques were developed to obtain α-expectation and two-sided (α, β) tolerance intervals for the two-way nested random effects model. This paper presents a simulation study to assess the statistical properties of these non-parametric techniques against classical and Bayesian techniques. Results show that the non-parametric techniques provided relatively small intervals, and generally retain the nominal content and confidence levels, regardless of the underlying distribution.

Backtesting A One-Step Ahead Density Predictions of Value-at-Risk

Katleho Makatjane

Basetsana Consultants, Vereeniging, South Africa

Corresponding author: katlehomakatjane@basetsana.co.za

This study estimates a method for backtesting density forecasts that are based on a weighted threshold of a continuously ranked probability score. The weighting emphasizes regions of interest, such as the tails or the center of a variable's range, while retaining propriety, as opposed to a recently developed weighted likelihood ratio test, which can be hedged, is emphasized in this study. Threshold-based decompositions of a continuously ranked probability score is illustrated and prompt insights into strengths and deficiencies of a forecasting method. A time-varying model with combined extreme values that are based on the score function of a predictive density at time is estimated. The mechanism to update the parameters of this model overtime is based on the scaled score of a likelihood function. The developed backtesting procedure revealed that the estimated generalized autoregressive score-generalized extreme value distribution (GAS-GEVD) model with a skewed student-t distribution has the best prediction performance in forecasting the value-at-risk (VaR) and Expected shortfall. Extension of this non-stationary distribution in literature is quite complicated since it requires specifications not only on how the usual Bayesian parameters change over time but also those with bulk distribution components. This implies that the combination of a stochastic econometric model with extreme value theory (EVT) procedures provides a robust basis necessary for the statistical backtesting density predictions for Value-at-risk (VaR).

Differential Networks as Association Change Detection Tools

Ricardo Daniel Marques Saldogo

University of Pretoria, Pretoria, South Africa

Corresponding author: ricardodansalgado@gmail.com

Graphical network modeling is a new modern technique to account for the connection and dependencies among the features in the genomic, biological, and data science fields. Through graph vertices, we measure the precision estimation of a graphical network. A more complex structure, called differential networks (DNs), involving the difference between two undirected graphical model(s) precisions, is handy for measuring the variations and changes in the connectivity. With the aid of DN, we can realize and detect the association change between two models and significantly judge the change employing some metrics. DN analysis is attributed mainly to their ability to effectively represent the relationships between factors of complex systems over time or various experimental conditions. However, a differential network is not easily calculated due to thresholding problems, and in the high-dimensional settings common within biological sciences, they must be estimated. Thus, there is a demand to explore different approaches for accurately and efficiently evaluating a differential network to derive and develop a state-of-the-art framework that will form the standard for differential network estimation in the future. We accurately capture the hidden correlation pattern among the features in genomics, biological, and data science fields using DNs.

Model comparison for Bayesian ALT models

Neill Smit[1] and Lizanne Raubenheimer[2]

[1] *North-West University, Vanderbijlpark, South Africa*

[2] *Rhodes University, Makhanda, South Africa*

Corresponding author: neillsmit1@gmail.com

criterion. An alternative and more formal approach is to use Bayes factors to compare models. However, Bayesian accelerated life testing models with more than one stressor often have mathematically intractable posterior distributions and Markov chain Monte Carlo methods are employed to obtain posterior samples to base inference on. The computation of the marginal likelihood, needed to calculate the Bayes factors, is challenging when working with such complex models. In this paper, methods for approximating the marginal likelihood are investigated. The focus is on methods that do not further complicate the Markov chain Monte Carlo algorithm, and where the marginal likelihood can be easily estimated from the posterior samples generated. The deviance information criterion and Bayes factors are compared in an application to a real data set, within the accelerated life testing paradigm for dual-stress models.

Variational Autoencoders to Enhance Classification Accuracy in Photovoltaic Fault Detection

Edward James Westraadt, Chantelle Clohessy and Warren Brettenny

Nelson Mandela University, Gqeberha, South Africa

Corresponding author: s215052064@mandela.ac.za

The detection and identification of faults arising in photovoltaic (PV) modules is essential to maintain the efficiency, safety and reliability of PV systems. If these faults are not detected, systems may experience extreme energy loss, shutdowns, safety concerns and, inevitably, financial losses (Garoudja et al, 2017). In larger PV installations, checking or testing individual PV modules may be extremely time and labour intensive, resulting in higher maintenance costs. Simplifying such a task to an automated, or semi-automated, process would benefit such large PV systems immensely. This study assesses three convolutional neural networks (CNNs) for the detection and classification of faults in PV modules. This deep learning approach follows on the work of Dunderdale et al (2020) of Nelson Mandela University, in an effort to find the most accurate technique for the detection and classification of PV faults using infrared thermal imagery. Variational Autoencoders (VAEs) were used to inflate the small dataset, and results were collected using the three CNN architectures, namely: InceptionV3, ResNet50 and Xception. Results obtained compare favourably to those of Dunderdale et al (2020), and further expand the knowledge in such a research area, introducing the use of VAEs to increase smaller dataset sizes, especially for the robust training of the CNNs.

Multivariate Analysis of Medical Schemes Marketing Expenditure

Michael Willie

Council For Medical Schemes, Pretoria, South Africa

Corresponding author: m.willie@medicalschemes.co.za

Marketing strategies are viewed as an investment in many corporate entities, and these are often used as tools to maximise the shareholders' return. Studies have shown that poor marketing strategies can lead to poor organizational performance. This study's primary objective of this paper was to assess the extent to which factors affect marketing activities and expenditure impact on scheme performance. The study entailed a univariate analysis of factors that affect marketing activities and expenditure and their impact on scheme performance. Multivariate analysis was employed to compare marketing expenditure in medical schemes. The analysis covered marketing expenditure data from 54 medial schemes, a convenience sampling frame. The number of beneficiaries that will be accounted for in the sample will be 68% and 65% of all beneficiaries and the marketing expenditure reported in 2019, respectively. The number of benefit options also attracted a higher marketing expense for medical schemes with more than four benefit options attracted more elevated levels of marketing expenditure. The business operating model was also a critical factor in marketing expense. Marketing expenditure for schemes with an insourced operating model (n=8) spends more on marketing activities than those with an outsourced business operating model (46). This study found some evidence of key factors that impact marketing expenditure, however, these were determinants of organisational performance, both in market share and membership and financial performance.