SASA2024 Oral Presentation Abstracts

Table of Contents

65 Spatial linear network Voronoi analysis to quantify accessibility of police stations in SA
Mr Arthur Antonio[1], Prof Inger Fabris-Rotelli[1], Dr Renate Thiede[1], Dr Rene Stander[1]
[1]University of Pretoria, Pretoria, South Africa

160 Estimating the incubation period of COVID-19
Mr Lebogang Baloi[1], Dr Najmeh Nakhaei Rad[1], Prof Din Chen[1]
[1]University of Pretoria, Pretoria, South Africa

74 Valuation of life insurance business with deep neural networks
Mr Jan Blomerus[1]
[1]University of the Free State, Bloemfontein, South Africa

141 Unravelling the dynamics of GGE biplots as visualisation tool to interpret and present agricultural trials
Dr Mardé Booyse[1], Dr Zelda Bijzet[1]
[1]Agricultural Research Council, Stellenbosch, South Africa

108 Multivariate stratified sampling allocation
Ms Georgi Borros[1], Dr Sebnem Er[1], Mr Sulaiman Salau[1]
[1]University of Cape Town, Cape Town, South Africa

128 Quantifying the directional relationship between natural hydrogen depressions and fault lines in Mpumalanga, South Africa
Mr Calvin Jens Botha[1], Dr Ansie Smit[2], Prof Inger Fabris-Rotelli[1], Dr Najmeh Nakhaei Rad[1], Prof Adam John Bumby[2], Dr Brenda Otieno Mac'Oduol[1]
[1]Department of Statistics, University of Pretoria, Pretoria, South Africa, [2]Department of Geology, University of Pretoria, Pretoria, South Africa

152 New classes of tests for the Weibull distribution in the presence of random right censoring
Dr Elzanie Bothma[1], Prof James Samuel Allison[1], Prof Izak Jacobus Hennig Visagie[1]
[1]North-West University, Potchefstroom, South Africa

155 Network analysis and disruption simulation of a South African cash-in-transit criminal network
Ms Annie Kok[1], Mr Stefan Britz[1]
[1]University of Cape Town, Cape Town, South Africa

168 bipl5: An R package for reactive calibrated axes biplots
Mr Ruan Buys[1], Ms Delia Sandilands[1], Prof Sugnet Lubbe[1], Dr Carel Johannes van der Merwe[1]
[1]Stellenbosch University, Stellenbosch, South Africa

64 A consolidated approach to linear mixed models with factors having both fixed and random levels
Dr Lyson Chaka[1], Prof Peter M. Njuho[2], Prof Hans-Peter Piepho[3]
[1]University of Pretoria, Pretoria, South Africa, [2]University of Zululand, Emphangeni, South Africa, [3]University of Hohenheim, Stuttgart, Europe

97 On precedence tests with double sampling
Dr Niladri Chakraborty[1], Prof Narayanaswamy Balakrishnan[2], Prof Maxim Finkelstein[1]
[1]University of the Free State, Bloemfontein, South Africa, [2]McMaster University, Hamilton, Canada

137 Modelling divest South African stock prices using mixture distribution
Prof Martin Chanza[1,2,4], Dr Modisane Seitshiro[3,4]
[1]Technology Enhanced Learning and Innovative Education and Training in South Africa (TELIT-SA), North-West University, Vanderbijlpark, South Africa, [2]Department of Statistics and Operations Research, North-West University, Mafikeng, South Africa, [3]Centre for BMI, North-West University, Potchefstroom, South Africa, [4]National Institute for Theoretical and Computational Sciences (NITheCS), South Africa

40 Experimental designs for estimating non-linear models in mixture variables
Prof Roelof Coetzer[1]
[1]North-West University, Potchefstroom, South Africa

101 Seasonal catchment areas using an attribute based fuzzy lattice data structure
Mrs Michelle de Klerk[1], Prof Inger Fabris-Rotelli[1]
[1]University of Pretoria, Pretoria, South Africa

167 Shewhart X control charts for monitoring the mean of autocorrelated AR(1) data
Dr MD Diko[1], Mr Tiisetso Molele[1], Mr Ntlhari Mabunda[1], Mr Matsebe Mabotha[1]
[1]University of the Free State, Bloemfontein, South Africa

136 Design of spatial capture recapture (SCR) surveys for stratified populations
Dr Greg Distiller[1,2], Associate Prof Ian Durbach[1,2], Ms Anita Wilkinson[3]
[1]Department of Statistical Sciences, University of Cape Town, Cape Town, South Africa, [2]Centre for Statistics in Ecology, the Environment, and Conservation, Cape Town, South Africa, [3]The Cape Leopard Trust, Cape Town, South Africa

50 Understanding the impact of parameter estimates on model performance
Lindani Dube[1,3], Tatenda Shoko[2], Tanja Verster[1,3]
[1]North-West University Centre for BMI, Potchefstroom, South Africa, [2]AIMS, Stellenbosch, South Africa, [3]NITheCS, Stellenbosch, South Africa

164 Exploratory spatial analysis of early grade reading data in KwaZulu-Natal
Mr Joshua Engelbrecht[1]
[1]JET Education Services, Johannesburg, South Africa

21 Trends in quantity and demographic composition of statistics graduates at South African universities, 1986-2022
Dr Thomas Farrar[1]
[1]Cape Peninsula University of Technology, Bellville, South Africa

46 Development and implementation of fictional narratives for enriched teaching of university-level statistics
Prof Johan Ferreira[1], Dr Seite Makgai[1]
[1]University of Pretoria, Pretoria, South Africa

135 Logratio analysis (LRA) and compositional biplots of milk fatty acids
Dr Susan Laurens[1], Dr Raeesa Ganey[2]
[1]Heineken Beverages, Stellenbosch, South Africa, [2]University of the Witwatersrand, Johannesburg, South Africa

62 Mapping linguistic beauty: biplot analysis of 228 world language patterns
Dr Raeesa Ganey[1,2], Dr Johané Nienkemper-Swanepoel[2,3]
[1]School of Statistics and Actuarial Science, University of Witwatersrand, Johannesburg, South Africa, [2]Centre for Multi-Dimensional Data Visualisation (MuViSU), Stellenbosch University, Stellenbosch, South Africa, [3]Department of Statistics and Actuarial Science, Stellenbosch University, Stellenbosch, South Africa

130 Type I multivariate Pólya-Aeppli distributions with applications
Ms Claire Geldenhuys[1], Dr René Ehlers[1], Prof Andriette Bekker[1]
[1]Department of Statistics, University of Pretoria, South Africa

57 Dynamic prediction and standard prediction models for type 2 diabetic individuals in the Western Cape
Mr Frissiano Honwana[1], Associate Prof F Gumedze[1], Prof L Myer[1], Dr J Rusch[2]
[1]University of Cape Town, Cape Town, South Africa, [2]National Health Laboratory Service, Cape Town, South Africa

150 Identifying differences between batters in Twenty20 cricket using principal component analysis and biplots
Mr Cameron Howe-Dreyer[1], Dr Paul J van Staden[1], Prof Inger N Fabris-Rotelli[1]
[1]Department of Statistics, University of Pretoria, Pretoria, South Africa

170 Seasonal volatility patterns in SAFEX grain futures: analysing environmental and supply-side risks
Dr Ayesha Sayed[1], Associate Prof Chun-Sung Huang[1]
[1]Department of Finance and Tax, University of Cape Town, Cape Town, South Africa

126 Stakeholder focused explainable artificial intelligence
Ms Gandhi Jafta[1], Prof Inger Fabris-Rotelli[1], Prof Emma Ruttkamp-Bloem[1], Dr Iketle Maharela[1]
[1]University of Pretoria, Pretoria, South Africa

93 Modelling toroidal data for representation and analysis of protein dihedral angles
Mr Claudio Jardim[1], Prof Inger Fabris-Rotelli[1], Dr Alta de Waal[1], Dr Najmeh Nakhaei Rad[1]
[1]University of Pretoria, Pretoria, South Africa

82 The impact of environmental shocks due to climate change on intimate partner violence: a SEM
Ms Esme Jordaan[1,2], Ms Jenevieve Mannell[3,4]
[1]Biostatistics Research Unit, South African Medical Research Council, Parow, South Africa, [2]Statistics and Population Studies, University of the Western Cape, Cape Town, South Africa, [3]Institute for Global Health, UCL, London, UK, [4]National University of Samoa, Samoa

48 Emailed publication invitations received by biostatisticians: academically sound versus potentially predatory journals
Prof Gina Joubert[1], Dr Omololu Aluko[1]
[1]University of the Free State, Bloemfontein, South Africa

106 A stochastic modelling of South African COVID-19 mortality, new infections and vaccination dynamics
Mr Malandala Kajingulu[1], Prof Edmore Ranganai[1]
[1]University of South Africa, Johannesburg, South Africa

27 Non-parametric methods for forecasting South African maize and wheat prices
Ms. Emelia Kammies[1]
[1]Sol Plaatje University, Kimberley, Northern Cape, South Africa

10 Virtual screening of plants and compounds against various disease targets using machine learning
Mr Alexander Kelbrick[1], Dr Najmeh Nakhaei Rad[1], Dr Paul van Staden[1], Prof Vinesh Maharaj[2]
[1]University of Pretoria, Pretoria, South Africa, [2]Department of Chemistry, University of Pretoria

71 A discrete-time competing risk analysis of students' academic behaviour: cause-specific and subdistribution hazards approach
Dr Lionel Establet Kemda[1]
[1]Durban University of Technology, Durban, South Africa

22 Taking data science collaboration to new heights in a study to better understand perceived versus actual digital behaviour
Mr Fallo Happy Khanye[1,2], Prof Renette Blignaut[1], Dr Julia Keddie[1]
[1]University of the Western Cape, Bellville, South Africa, [2]Ghent University, Ghent, Belgium

66 Entropy penalised self-paced learning
Mr Andre Ruben Kleynhans[1], Prof Frans Kanfer[1], Prof Sollie Millard[1]
[1]University of Pretoria, Pretoria, South Africa

144 Economic recession prediction using modified gradient boosting and principal component neural network algorithms
Mr Anuroop Krishnannair[1], Dr Najmeh Nakhaei Rad[1]
[1]University of Pretoria, Pretoria, South Africa

133 Bayesian approach to the estimation of asymptotic dependence and independence in joint tails
Mr Nicholas Kwaramba[1], Prof Andrehette Verster[1]
[1]University of the Free State, Bloemfontein, South Africa

73 A combined point process for better-than-minimal, minimal, and worse-than-minimal repairs
Miss Amy Langston[1], Prof Maxim Finkelstein[2,3], Prof Ji Hwan Cha[4]
[1]Rhodes University, Makhanda, South Africa, [2]University of the Free State, Bloemfontein, South Africa, [3]University of Strathclyde, Glasgow, Scotland, [4]Ewha Womans University, Seoul, Republic of Korea

124 Timeseries PCA biplots
Prof Sugnet Lubbe[1]
[1]Stellenbosch University, Stellenbosch, South Africa

84 Generalising the molecular speed distribution of Maxwell
Prof Iain MacDonald[1], Dr Etienne Pienaar[1]
[1]University of Cape Town, Cape Town, South Africa

140 Identifying contributing factors to profile non-completing students in the faculty of natural sciences
Mr Edwin Mahlangu[1], Ms Xabsa Mohumed[1], Ms Kesia Phigeland[1], Dr Humphrey Brydon[1], Ms Khadija Parker[1], Prof Renette Blignaut[1]
[1]University of the Western Cape, Bellville, South Africa

103 Enhanced point pattern analysis on nonconvex spatial domains
Mr Kabelo Mahloromela[1], Prof Inger Fabris-Rotelli[1]
[1]University of Pretoria, Pretoria, Gauteng, South Africa

12 Extreme value dependence analysis to bitcoin/us dollar and South African rand/us dollar exchange rates
Dr Katleho Makatjane[1], Mr Lethlogononolo Mosanawe[1], Dr Claris Shoko[1]
[1]University of Botswana, Gaborone, Botswana
80 Distribution-free generalised EWMA control charts using two-sample tests with application in froth flotation process
Miss Palesa Makena[1], Dr Majika Jean Claude Malela[1]
[1]University of Pretoria, Pretoria, South Africa

165 Quantifying how fast South Africa's new car sales recovered from the COVID-19 pandemic using time series intervention analysis
Dr Tendai Makoni[1], Prof Delson Chikobvu[1]
[1]University of the Free State, Bloemfontein, South Africa

49 A rank-based EWMA TBEA control chart
Dr Majika Jean Claude Malela[1], Prof Fernanda Otilia Figueiredo[2], Prof Philippe Castagliola[3]
[1]University of Pretoria, Pretoria, South Africa, [2]University of Porto, Porto, Portugal, [3]University of Nantes, Nantes, France

107 Multivariate Bayesian small area estimation of health statistics indicators
Prof Samuel Manda[1]
[1]University of Pretoria, Pretoria, South Africa

134 Trend analysis and determinants of violence against women in South Africa using VOCS 2013-2017 data
Mr Sonnyboy Manthata[1], Dr Lebogang Sesale[2], Prof Solly Seeletse[3]
[1]Sefako Makgatho University of Health Sciences, Pretoria, South Africa, [2]University of South Africa, Pretoria, South Africa, [3]Sefako Makgatho University of Health Sciences, Pretoria, South Africa

139 An analysis of new entrants in technical and vocational education and training colleges: 2022
Mr Sonnyboy Manthata[1], Ms Nthabiseng Tema[1], Ms Mmakgotso Ntsoane
[1]Department of Higher Education and Training, Pretoria, South Africa

18 Quantifying loss to the SA wholesale and retail industries using interrupted time series models
Mr Thabiso Masena[1], Mr Sandile Shongwe[1], Dr Ali Yeganeh[1]
[1]University of the Free State, Bloemfontein, South Africa

42 Application of extreme value theory to finance data
Mr Daniel Levy Mashilo[1]
[1]University of South Africa, Johannesburg, South Africa

161 The impact of clustering in randomised clinical trials: scoping review and comparative statistical analysis
Ms Mikateko Mazinu[1], Prof S Manda[2], Dr T Reddy[3]
[1]South African Medical Research Council, Tygerberg, South Africa, [2]University of Pretoria, Pretoria, South Africa, [3]South African Medical Research Council, Durban, South Africa

158 An analytical and empirical comparison of meta-analysis methods for individual participant binary data
Ms Abigail Mberi[1], Prof Samuel Manda[1]
[1]University of Pretoria, Pretoria, South Africa

39 Determination of predictors related to high blood pressure in South Africa using machine learning techniques
Dr Ruffin Mpiana Mutambayi[1], Mr Nhlonipho Mbhele[1], Mr Masimthembe Lala[1]
[1]University of Fort Hare, Alice, South Africa

52 Proportion and risk factors associated with 'never tested for HIV' amongst women in Tanzania
Dr Sizwe Mbona[1], Prof Retius Chifurira[1], Dr Bonginkosi Duncan Ndlovu[1]
[1]Durban University of Technology, Durban, South Africa

20 Application of joint modelling and longitudinal latent modelling to antiretroviral adherence monitoring
Mr Campbell Mcduling[1], Ms Lauren Jennings[2], Prof Catherine Orrell[2], Prof Francesca Little[1]
[1]Department of Statistics, University of Cape Town, Cape Town, South Africa, [2]Center for Adherence and Therapeutics, Desmond Tutu Health Foundation, Cape Town, South Africa

89 Logistic regression analysis to identify the determinants of concurrent sexual partnership among Kenyan women
Dr Tshaudi Motsima[1], Mr Banele Mdakane[1], Ms Thelma Maunye[1]
[1]Tshwane University of Technology, Pretoria, South Africa

79 Survival analysis of time-to-credit default in the presence of time-varying covariates
Mr Lusanda Mdhlalose[1]
[1]University of the Witwatersrand, Johannesburg, South Africa

15 Use of some important statistical methods in electrical energy generation and their applications
Dr Vincent Micali[1]
[1]Stats4buz (Pty) Ltd, Hout Bay, Cape Town, South Africa

85 Divergence-based approach in bivariate tail dependence coefficient estimation
Dr Richard Minkah[1,3], Prof Abhik Ghosh[2], Prof Tertius de Wet[3]
[1]University of Ghana, Accra, Ghana, [2]Indian Statistical Institute, Kolkata, India, [3]Stellenbosch University, Stellenbosch, South Africa

78 Analysing exercise-associated muscle cramping in ultramarathon runners using predictive modelling
Ms Xabsa Ahmed Mohumed[1], Dr Retha Luus[1], Ms Tayla Wannenberg[1], Mrs Esme Jordaan[2]
[1]University of the Western Cape, Bellville, South Africa, [2]South African Medical Research Council, Parow, South Africa

94 Comparing the power of multivariate test statistics for three-factor interaction in a 3-way contingency table
Ms PB Mokoena[1]
[1]University of South Africa, Florida, South Africa

9 Estimation of covariance function of a stationary ARMA process
Dr Wessel Moolman[1]
[1]Akademia, Centurion / Die Hoewes, South Africa

81 Predictors of emotional and physical abuse towards Kenyan men: a logistic regression analysis
Dr Tshaudi Motsima[1]
[1]Tshwane University of Technology, Pretoria, South Africa

95 An illustration of gender differential item functioning analysis in mathematics from national benchmark tests
Mrs Precious Mudavanhu[1]
[1]University of Cape Town, Cape Town, South Africa

38 Analysis of predictors related to diagnosis of hypertension correlated with heart attacks in South Africa
Dr Ruffin Mpiana Mutambayi[1], Mrs Natalie Benschop[2], Prof Retius Chifurira[2]
[1]University of Fort Hare, Alice, South Africa, [2]University of KwaZulu-Natal, Durban, South Africa

173 Distributions of wet and dry spells
Ms Nothabo Ndebele[1]
[1]University of the Witwatersrand, Johannesburg, South Africa

59 A nonparametric estimation of cumulative incidence functions in the presence of cured subjects
Dr Bonginkosi Duncan Ndlovu[1], Dr Sizwe Vincent Mbona[1]
[1]Department of Statistics, Durban University of Technology, Durban, South Africa

14 The GARCH-EVT – Gumbel copula approach to quantifying portfolio diversification effects
Mr Thabani Ndlovu[1], Prof Delson Chikobvu[1]
[1]University of the Free State, Bloemfontein, South Africa

125 Designing an optimal survey sample with predetermined sample sizes for subgroups
Dr Ariane Neethling[1], Mr Francois Neethling[2]
[1]Independent Statistical Consultant, Bloemfontein, South Africa, [2]Independent Statistical Consultant, Cape Town, South Africa

67 Optimal grid selection in spatial statistics
Ms Jamie-Lee Nel[1], Dr René Stander[1], Mr Kabelo Mahloromela[1], Prof Inger N Fabris-Rotelli[1]
[1]University of Pretoria, Pretoria, South Africa

163 Parametric analysis of multistate survival modelling for birth parity transitions in rural South Africa
Ms Thambeleni Portia Nevhungoni[1], Dr Tarylee Reddy[1], Prof Samuel Manda[2], Prof Din Chen[2]
[1]South African Medical Research Council, Pretoria, South Africa, [2]University of Pretoria, Pretoria, South Africa

88 Understanding macroeconomic factors' influence on South African maize production and food security: VECM analysis
Ms Cynthia Boitumelo Ngwane[1,2], Mr Kajingulu Malandala[1]
[1]University of South Africa, Florida, South Africa, [2]Agricultural Research Council, Pretoria, South Africa

98 Application of marginal theory for variable selection in partially linear models
Dr Mina Norouzirad[1], Dr Ricardo Moura[1], Prof Mohammad Arashi[2], Dr Filipe Marques[3]
[1]Center for Mathematics and Applications (NovaMath), NOVA School of Science and Technology (NOVA FCT), Caparica, Portugal, [2]Department of Statistics, Faculty of Mathematical Sciences, Ferdowsi University of Mashhad, Mashhad, Iran, [3]Center for Mathematics and Applications (NOVA Math) and Department of Mathematics, NOVA School of Science and Technology (NOVA FCT), Caparica, Portugal

19 Profiling students at risk of dropout at a university in South Africa
Dr Piet Ntema[1]
[1]University of Limpopo, Polokwane, South Africa

154 Statistical data analysis of multidimensional binary data using chi-square tests and correspondence analysis
Ms Nombasa Ntushelo[1], Dr Itumeleng Matle[1], Mr Thabo Nkabinde[2]
[1]Agricultural Research Council, Stellenbosch, South Africa, [2]Agrispace, Cape Town, South Africa

90 The analysis of the cosmological parameters using maximum likelihood estimator and chi-square
Miss Sinenhlanhla Nxumalo[1], Prof Syamala Krishnannair[1]
[1]University of Zululand, KwaDlangezwa, South Africa

69 Enhanced process monitoring: the bootstrapped cumulative sum- exponentially weighted moving average (BCUSUM-EWMA) control chart approach
Dr Braimah Joseph Odunayo[1], Prof Fabio Correa[1]
[1]University of the Free State, Bloemfontein, South Africa

44 Cytokine profiles as predictors of HIV incidence using machine learning survival models and statistical interpretable techniques
Ms Sarah Ogutu[1], Dr Mohanad Mohammed[1], Prof Henry Mwambi[1]
[1]University of KwaZulu-Natal, Pietermaritzburg, South Africa

122 Stochastic modelling on rainfall variability in northern Nigeria
Prof John Olaomi[1], Mr James Ehimony[1,2]
[1]Department of Statistics, University of South Africa, Florida, Johannesburg, South Africa,
[2]Department of Statistics, Kogi State Polytechnic, Lokoja, Nigeria

118 A contaminated negative binomial model for count health data
Mr Arno Otto[1]
[1]University of Pretoria, Pretoria, South Africa
104 Investigating the robustness of clustered point pattern simulation
Miss Amy Pieters[1], Dr Rene Stander[1], Prof Inger Fabris-Rotelli[1], Mr Kabelo Mahloromela[1], Dr Renate Thiede[1]
[1]University of Pretoria, Pretoria, South Africa

16 Clustering and classifying global food insecurity index and crop production using machine learning algorithms
Mr Jaden Pieterse[1]
[1]Sol Plaatje University, Kimberley, South Africa

43 The importance of specification of the deterministic components in the co-integration model: using data on employment costs and gross earnings to show the impact of model misspecification
Dr Sagaren Pillay[1]
[1]Statistics South Africa, Pretoria, South Africa

127 Soft clustering missing at random (MAR) data
Mr Jason Pillay[1]
[1]University of Pretoria, Pretoria, South Africa

53 Transferability of GANs-UNet model for informal road detection in underdeveloped areas
Miss Luandrie Potgieter[1], Prof Inger Fabris-Rotelli[1], Dr Renate Thiede[1]
[1]University of Pretoria, Hatfield, Gauteng

24 Break detection in high-dimensional panel data
Prof Marie Hušková[1], Prof Charl Pretorius[2]
[1]Charles University, Prague, Czech Republic, [2]Centre for BMI, North-West University, Potchefstroom, South Africa

23 GPAbin biplots for continuous data: a methodology for combining biplots of completed continuous data sets
Mr Mokgeseng Ramaisa[1], Dr Johané Nienkemper-Swanepoel[1]
[1]Department of Statistics and Actuarial Science, Centre for Multi-Dimensional Data Visualisation (MuViSU), Stellenbosch University, Stellenbosch, South Africa

143 Enhancing research guidance in Statistics supervision: adapting to the generative AI era
Dr Danielle Roberts[1], Prof Inger Fabris-Rotelli[2], Prof Sonali Das[2], Prof Michael von Maltitz[3], Dr Ansie Smit[2], Prof Daniel Maposa[4], Prof Fabio Correa[3]
[1]University of KwaZulu-Natal, Durban, South Africa, [2]University of Pretoria, Pretoria, South Africa, [3]University of the Free State, Bloemfontein, South Africa, [4]University of Limpopo, Polokwane, South Africa

55 Compositional biplot approaches
Mr Phuti Sebatjane[1,2], Prof Sugnet Lubbe[2], Prof Niël le Roux[2]
[1]Department of Statistics, University of South Africa, Pretoria, South Africa, [2]MuViSU, Department of Statistics and Actuarial Science, Stellenbosch University, Stellenbosch, South Africa

146 Bayesian prior elicitation for malaria modelling
Ms Makwelantle Asnath Sehlabana[1], Prof Daniel Maposa[2], Dr Alexender Boateng[3], Prof Sonali Das[4]
[1]University of Limpopo, Polokwane, South Africa, [2]University of Limpopo, Polokwane, South Africa, [3]Department of Mathematics and Computer Science, Modern College of Business and Science, Bawshar, Oman, [4]University of Pretoria, Pretoria, South Africa

105 Comparative analysis of the return level estimates based on block maxima and POT extreme value theory approaches
Ms Anna Seimela[1], Prof Daniel Maposa[1]
[1]University of Limpopo, Polokwane, South Africa

63 Enhancing financial market risk measures: a comparative analysis of long-memory GARCH-type models
Dr Modisane Seitshiro[1]
[1]North-West University, Potchefstroom, South Africa

54 Spatial dependency modelling of disjoint spatial areas - SAPRIN urban node analysis
Ms Ephent Selahle[1], Prof Inger Fabris-Rotelli[1], Mrs Nada Abdelatif[2]
[1]University of Pretoria, Pretoria, South Africa, [2]South African Medical Research Council, Cape Town, South Africa

110 Estimating disability rates in South African districts using area-level Poisson mixed models
Prof Yegnanew Shiferaw[1]
[1]Department of Statistics, University of Johannesburg, Johannesburg, South Africa

25 Data-driven approaches for predicting electricity demand
Prof Caston Sigauke[1]
[1]University of Venda, Thohoyandou, South Africa

132 Joint modelling for longitudinal and interval censored survival data
Dr Isaac Luwinga Singini[1], Prof Ding-Geng Chen[2], Associate Prof Freedom Gumedze[3]
[1]Biostatics Research Unit, South African Medical Research Council, Cape Town, South Africa, [2]Department of Statistics, University of Pretoria, Pretoria, South Africa, Pretoria, South Africa, [3]Statistical Sciences Department, University of Cape Town, Cape Town, South Africa

138 An assessment of the impact of spatial connectivity structures on spatial model fit: machine-learning approach
Dr Claris Siyamayambo[1], Dr Edith Phalane[1], Prof Refilwe Phaswana-Mafuya[1], Prof Inger Fabris-Rotelli[2]
[1]University of Johannesburg, Johannesburg, South Africa, [2]University of Pretoria, Pretoria, South Africa

129 Profile-likelihood based confidence intervals in earthquake hazard assessment models
Mr Siyamthanda Prusent[1], Dr Ansie Smit[2], Prof Inger Fabris-Rotelli[1], Dr Najmeh Nakhaei Rad[1], Dr Brenda Otieno Mac'Oduol[1]
[1]Department of Statistics, University of Pretoria, [2]Department of Geology, University of Pretoria

28 A threshold-search approximate Bayesian computation algorithm for parameter estimation
Dr Neill Smit[1]
[1]North-West University, Potchefstroom, South Africa

37 Goodness-of-fit tests with applications in risk modelling
Ms Leoni Snyman[1], Prof James Allison[1], Prof Jaco Visagie[1], Prof Simos Meintanis[2]
[1]North-West University, Potchefstroom, South Africa, [2]University of Athens, Athens, Greece

91 An improved test for the accuracy of spatial point pattern tests
Dr Rene Stander[1], Prof Inger Fabris-Rotelli[1], Prof Gregory Breetzke[1], Dr Jean-Pierre Stander[1]
[1]University of Pretoria, Pretoria, South Africa

7 Comparing the asymptotic relative efficiency of the CMP model with the negative binomial model
Dr Yuvraj Sunecher[1]
[1]University of Technology, Pointe Aux Sables, Mauritius

76 A generalised homogeneously weighted moving average scheme for monitoring the process mean
Mr Maonatlala Thanwane[1], Dr Majika Jean Claude Malela[1], Prof Frans Kanfer[1], Prof Kashinath Chatterjee[2]
[1]University of Pretoria, Pretoria, South Africa, [2]University of Augusta, Georgia, United States of America

58 A statistical exploration of the effect of road network structure on road-based accessibility
Dr Renate Thiede[1], Prof Inger Fabris-Rotelli[1], Prof Pravesh Debba[2], Prof Christopher Cleghorn[3]
[1]University of Pretoria, Pretoria, South Africa, [2]CSIR, Pretoria, South Africa, [3]University of the Witwatersrand, Johannesburg, South Africa

116 Prevalence and risk factors associated with HIV infection among pregnant antenatal attendees in Limpopo Province
Dr Oratilwe Penwell Mokoena[1], Mr Donald Tshabalala[1], Dr Thembelihle Sam Ntuli[1], Mr IT Boshomane
[1]Sefako Makgatho University

51 A quantile regression model for bounded longitudinal data and survival data
Dr Divan A Burger[1,2,3], Dr Sean van der Merwe[2], Dr Janet van Niekerk[4,3], Prof Emmanuel Lesaffre[5], Mr Antoine Pironet[6]
[1]Syneos Health, Bloemfontein, South Africa, [2]University of the Free State, Bloemfontein, South Africa, [3]University of Pretoria, Pretoria, South Africa, [4]King Abdullah University of Science and Technology, Thuwal, Kingdom of Saudi Arabia, [5]KU Leuven, Leuven, Belgium, [6]AARDEX Group, Liège, Belgium

166 A noncentral Poison-Lindley distribution contextualised in a process monitoring framework
Dr Ané van der Merwe[1], Prof Johan Ferreira[1]
[1]University of Pretoria, Pretoria, South Africa

102 From hotspot detection to accessibility: a spatial network analysis of informal settlements
Mr R van der Walt[1], Dr R.N. Thiede[1], Prof I.N. Fabris-Rotelli[1]
[1]University of Pretoria, Pretoria, South Africa

45 Mapping soil thickness by accounting for right-censored data with survival probabilities and machine learning
Dr Stephan van der Westhuizen[1,2,3], Prof G. B. M. Heuvelink[2,3], Dr D. P. Hofmeyr[4], Dr L Poggio[3], Dr M Nussbaum[5], Prof C Brungard[6]
[1]Stellenbosch University, Stellenbosch, South Africa, [2]Wageningen University, Wageningen, the Netherlands, [3]ISRIC - World Soil Information, Wageningen, the Netherlands, [4]Lancaster University, Lancaster, United Kingdom, [5]Utrecht University, Utrecht, the Netherlands, [6]New Mexico State University, Las Cruces, United States of America

169 An adjustive rating system for rugby union based on exponential smoothing
Dr Paul J van Staden[1], Mr Waldo Botha[1], Mr Carlo Geyer[1]
[1]Department of Statistics, University of Pretoria, Pretoria, South Africa

113 Bayesian variable selection for skew-normal models
Mr Arnold van Wyk[1], Prof Andriette Bekker[1], Prof Mohammad Arashi[2], Dr Janet van Niekerk[1,2]
[1]University of Pretoria, Pretoria, South Africa, [2]Ferdowsi University of Mashhad, Mashhad, Iran

61 Insights into the construction of alternative bivariate cardioid distributions
Mrs Delene van Wyk-de Ridder[2], Prof Johan Ferreira[1], Prof Andriëtte Bekker[1]
[1]University of Pretoria, Pretoria, South Africa, [2]University of Cape Town, Cape Town, South Africa

100 The road not taken: spatial network optimisation on South African informal settlements
Ms C Van Zyl[1], Dr R.N. Thiede[1], Prof I.N. Fabris-Rotelli[1]
[1]University of Pretoria, Pretoria, South Africa

83 Sample size calculations in diagnostic accuracy studies with frequentist and Bayesian approaches
Mrs Lizelle Venter[1], Prof Ding-Geng Chen[2], Prof Inger Fabris-Rotelli[1]
[1]University of Pretoria, Pretoria, South Africa, [2]Arizona State University, Phoenix, U.S.A.

114 Construction and analyses of complete diallel cross through partially balanced incomplete block designs
Mr Anteneh Yalew[1], Prof M.K. Sharma[2]
[1]University of the Witwatersrand, Johannesburg, South Africa, [2]Addis Ababa University, Addis Ababa, Ethiopia

99 Geospatial small area estimation of hemoglobin levels of women and children in official statistics
Dr Seyifemickael Amare Yilema[1,2], Dr Najmeh Nakhaei Rad[1], Prof Ding-Geng Chen[1,3]
[1]Department of Statistics, University of Pretoria, Pretoria, South Africa, [2]Debre Tabor University, Debre Tabor, Ethiopia, [3]College of Health Solution, Arizona State University, Phoenix, USA

26 Application of longitudinal multilevel zero-inflated Poisson regression in modelling infectious diseases among infants in Ethiopia
Mrs Bezalem Eshetu Yirdaw[1], Prof Legesse Kassa Debusho[1], Dr Aregash Samuel[2]
[1]University of South Africa, Johannesburg, South Africa, [2]Ethiopian Public Health Institute, Gulele Sub City, Ethiopia

41 The future of programming in the age of GenAI
Mr Andre Zitzke[1]
[1]SAS, Johannesburg, South Africa

70

Exploring the bidirectional pathway between intimate partner violence and depression from a cluster randomised trial

Mrs Nada Abdelatif[1], Dr Esnat Chirwa[1], Dr Andrew Gibbs[2], Prof Samuel Manda[3]
[1]South African Medical Research Council, Cape Town, South Africa, [2]University of Exeter, Exeter, England, [3]University of Pretoria, Pretoria, South Africa

Intimate partner violence (IPV) predominately affects women and involves physical, sexual, emotional or psychological abuse by an intimate partner. IPV affects women of different cultures, cuts across geographical boundaries and settings, with approximately one third of women worldwide having experienced at least one form of violence. Furthermore, it has been shown that there is a strong association between IPV and mental health such as depression and depressive symptoms. Women who experience IPV are at higher risk of experiencing PTSD, anxiety disorders and suicidal ideation; but it has also been shown that those with depressive symptoms are at greater risk of being victims of IPV. To explore this bi-directional relationship between IPV and depression, women's experience of IPV and depression was analysed from a longitudinal cluster randomised trial conducted in South Africa. This was done using a bivariate probit model to jointly model the two outcomes and then by using a longitudinal structural equation model (SEM) to explore the bi-directional relationship between IPV and depression. A high correlation was found between IPV and depression. The SEM showed that previous experience of IPV and depression seem to increase depression in women.

151

An R Shiny application for optimising mixed medley team selection in Masters swimming

Miss San-Mari Ackerman[1], Dr Paul J van Staden[1], Prof Inger N Fabris-Rotelli[1]
[1]Department of Statistics, University of Pretoria, Pretoria, South Africa

This study investigates the optimisation of team selection dynamics in Masters swimming, focusing on mixed medley relay teams categorised by age. Through mathematical modelling, we incorporate stroke-specific constraints using Integer Linear Programming (ILP) to enhance performance. Freestyle, the most common stroke, shows the fastest times, while breaststroke and butterfly exhibit more significant performance declines, particularly for aging female swimmers. We developed an R Shiny application to streamline team composition by allowing users to input swimmer data and automatically generate optimal team configurations. This tool supports data-driven decision-making for swimming clubs and could serve as a template for team selection in other sports. Future research may expand the dataset and app functionalities to account for variables such as training history and competition rules, improving its applicability across various contexts.

77

The past, present and future of visualising sentiments

Miss Zoë-Mae Adams[1,2], Dr Johané Nienkemper-Swanepoel[1,2]
[1]Centre for Multi-dimensional Data Visualisation, Stellenbosch, South Africa, [2]Stellenbosch University, Stellenbosch, South Africa

Text visualisation aids in exploring and understanding the content of unfamiliar complex multivariate text data, where the graphical representation can either focus on visualising the raw text or the results of text mining techniques. The broad field of text visualisation has seen comprehensive research since the 1990s and entails techniques that summarise text content, display similarity or sentiments, and aid in big data exploration. However, literature on visualisations specifically aimed at representing sentiment visualisation, is limited. In cases where we are trying to gauge sentiment from the text data, sentiment visualisation can help provide a quick and intuitive understanding of the overall tone or emotions present in the text. In this presentation an extensive review of available methodologies with applications will be given to highlight the advances that have been made in recent years. Furthermore, it will expose the possibilities for further development in this field. This presentation therefore examines the evolution of sentiment visualisation.

171

Modelling determinants of contraceptive use among women in Nigeria using a hybrid ensemble approach

Dr Rotimi Felix Afolabi[1], Mr Oluwafemi Christopher Enabor[1], Prof Ayo Stephen Adebowale[1,2], Prof Martin E. Palamuleni[2]

[1]Department of Epidemiology and Medical Statistics, College of Medicine, University of Ibadan, Ibadan, [2]Population Studies and Demography Programme, North-West University, Mafikeng, South Africa

Low contraceptive use among women of childbearing age in Nigeria is a critical public health issue, contributing to overpopulation and associated poor maternal and child health outcomes. While existing research on predictors of contraceptive use often employs traditional methods, studies utilizing machine learning algorithms remain scarce. This study aimed to examine trends and model determinants of contraceptive use among Nigerian women through a hybrid ensemble approach. Utilizing a cross-sectional design, we analysed data from four consecutive rounds of the Nigeria Demographic and Health Survey (2003,2008,2013, and 2018). We employed decision-tree-based ensemble machine learning algorithms, specifically Random Forest and eXtreme Gradient Boosting, to model predictors of contraceptive use and compare their predictive performances with a standard Decision Tree. The prevalence of contraceptive use was 14.2%, ranging from 13.2% in 2003 to 16.0% in 2013. Overall, the prevalence of contraceptive use increased by 2.3% over the studied period. The Random Forest ensemble algorithm demonstrated the highest accuracy (95%) in predicting contraceptive usage among women in Nigeria. Key predictors identified included age, marital status, education level, wealth index, parity, place of residence, and partner's educational attainment. The findings underscore the effectiveness of the hybrid ensemble machine learning model, particularly the Random Forest algorithm, in accurately modelling and predicting contraceptive use. These insights highlight the critical factors that influence contraceptive use in Nigeria and underscore the urgent need for targeted interventions to enhance access to and utilization of family planning services, particularly for women with lower educational attainment and individuals within the lowest wealth index.

172

Decomposing factors influencing teenage pregnancy and motherhood in Nigeria, 2003 - 2018

Dr Rotimi Felix Afolabi[1], Dr Mobolaji M. Salawu[1], Prof Ayo Stephen Adebowale[1], Prof Martin E. Palamuleni[2]
[1]University of Ibadan, Ibadan, Nigeria, [2]North-West University, Mafikeng, South Africa

Nigeria is among the countries with a high burden of Teenage Pregnancy and Motherhood (TPM) in sub-Saharan Africa, with enormous adverse effect on young girls. This study aimed to assess trends, changes, and determinants of TPM shifts in Nigeria over the last two decades. Extracted from four consecutive rounds (2003, 2008, 2013, and 2018) of Nigeria Demographic and Health Survey datasets, data on women aged 20-49 years who had ever been pregnant, and reported at least one childbirth, pregnancy termination, or stillbirth before attaining age 20 were analysed using trend and multivariate decomposition analyses at a 5% significance level. The prevalence of TPM was 56.1%, ranging from 64.7% in 2003 to 55.7% in 2018. Overall, the prevalence of TPM decreased significantly by 10.7% over the studied period ($p<0.001$). The change was due to a composite of a positive significant effect of the net compositional change (126%) and a negative effect of the net behavioural change (26%). The identified significant drivers of shift in TPM due to changes in the composition of women included current age, educational level, timing of marriage, and region of residence. Due to the change in behaviours, TPM reduced by 20% among South-South residents compared with their North-Central counterparts. However, TPM increased by 260% among teens who had their first sexual initiation. The TPM prevalence remained high in Nigeria, though a decreasing trend was observed within the studied period. Government and other stakeholders should focus pragmatic interventions on the identified drivers of TPM change over the last two decades in their efforts to alleviate TPM in Nigeria.

109

Heteroscedastic accelerated failure time model for length-biased right-censored data

Dr Mahboubeh Akbari Lakeh[1], Dr Najmeh Nakhaei Rad[1], Prof Ding-Geng Chen[1,2]
[1]University of Pretoria, Pretoria, South Africa, [2]College of Health Solutions, Arizona State University, United States

Length-biased data arise when the probability of observing a subject in a sample is proportional to its corresponding value. This phenomenon has been widely recognised in various fields, including economics, industrial reliability, applications in etiology, and studies related to epidemiology, genetics, and cancer screening. The assessment of the relationship between risk factors and survival time in the presence of biased data, particularly length-biased right-censored (LBRC) data, has long been a statistical challenge. Since the structure of observed length-biased data differs from that of the target population, using traditional methods to estimate covariate effects based on the observed length-biased data is inappropriate. In this paper, we study the estimation of covariate effects in a heteroscedastic accelerated failure time model while the observations are subject to LBRC. The asymptotic properties including consistency and weak convergence of the estimates is derived. A simulation study is carried out to evaluate and compare the performance of the proposed procedures with the method that ignores heteroscedasticity. Finally, the procedures are illustrated by modelling the regression parameter for a set of real data.

156

A theoretical framework for correcting misspecification in geo-experiment ad campaigns

Dr Iman Al Hasani[1]
[1]Sultan Qaboos University, Alkoud, Oman

The study examines the potential impact of hidden heterogeneity in estimating the effectiveness of geo-experiment advertising campaigns. The problem arises from the fact that the campaign effect is usually estimated from a known misspecified model due to unobserved covariates. In this study, a theoretical framework is developed to correct the potential misleading inferences from the misspecified models. Model validation results suggest that the proposed theoretical framework is consistent, gets better with reliable design strategies and saves having to do expensive Monte Carlo simulations.

65

Spatial linear network Voronoi analysis to quantify accessibility of police stations in SA

Mr Arthur Antonio[1], Prof Inger Fabris-Rotelli[1], Dr Renate Thiede[1], Dr Rene Stander[1]
[1]University of Pretoria, Pretoria, South Africa

The overlap between current police precinct boundaries and theoretically optimal ones, derived from Voronoi diagrams using Euclidean and network distances, is quantified. Spatial similarity measures are applied to assess how boundary overlap impacts police station accessibility, with the hypothesis that reduced overlap leads to decreased access. The potential correlation between boundary placement and crime rates is also explored, suggesting that less accessible precincts may experience higher crime levels. By examining these dynamics, the effectiveness of current precinct boundaries and their potential influence on crime and public safety are evaluated.

160

Estimating the incubation period of COVID-19

Mr Lebogang Baloi[1], Doctor Najmeh Nakhaei Rad[1], Prof Din Chen[1]
[1]University of Pretoria, Pretoria, South Africa

The COVID-19 pandemic has highlighted the importance of accurately estimating the incubation period and generation time of infectious diseases. These parameters are crucial for effective epidemiological modelling and public health decision-making. The incubation period, defined as the interval between infection and symptom onset, is vital for determining optimal quarantine durations. Generation time, which is the period between the infection of a primary case and the occurrence of secondary cases, influences estimates of the basic reproduction number.

Commonly, the incubation distribution is estimated using contact-tracing-based methods. However, these methods are highly dependent on individuals' assessments of possible exposure dates, which can lead to significant errors. Alternative interval censoring-based methods can handle large datasets but may suffer from biased sampling. Observed serial intervals are often used to estimate the generation time distribution, but if the disease is infectious during the incubation period, these estimates can be biased.

In this research, we analyse a publicly available real dataset consisting of departure times from Wuhan and the onset of COVID-19 symptoms for 1,211 passengers. We make use of the incubation period as the interarrival time, and the duration between departure and symptom onset as a mixture of forward time and interarrival time with censored intervals. The incubation distribution is estimated using renewal process theory and interval censoring with a mixture distribution.

As a novel contribution, we derive that the incubation time follows the generalised gamma distribution and the generalised beta distribution of the second kind, which outperform existing models in the literature which assumed to be gamma and Weibull distributions. Consequently, a model selection procedure is examined with likelihood ratio statistics to confirm the superiority of these extended distributions. Additionally, a consistent estimator for the generation time distribution is obtained using the incubation period and serial intervals for incubation-infectious diseases.

103

Enhanced point pattern analysis on nonconvex spatial domains

Mr Kabelo Mahloromela[1], Prof Inger Fabris-Rotelli[1]
[1]University of Pretoria, Pretoria, Gauteng, South Africa

The analysis of point pattern data is done to expand the basic understanding of the properties of the underlying point process that generated the data. These properties are typically estimated using density and distance-based statistics which rely on the specification of the spatial domain. The spatial domain on which points are observed and the distance metric used to quantify proximity between the points thus play an important role in analysing point patterns. Convex windows with the Euclidean distance are conventionally used. This choice of window and distance measure, however, is not representative when used on spatial domains that are nonconvex and constrain points within them. Herein, we develop methodology to support the analysis of point pattern data on nonconvex spatial domains.

74

Valuation of life insurance business with deep neural networks

Mr Jan Blomerus[1]
[1]University of the Free State, Bloemfontein, South Africa

The life insurance industry relies on actuarial methods to determine the values of policy portfolios. These methods are time-consuming, complex, prone to errors, difficult to audit and use costly infrastructures to produce.

In this study, a dataset of a commercial European life insurer is used to perform traditional actuarial pricing and valuation exercises. I then applied a deep neural network to calibrate the results to those of the traditional methods.

The results of my study indicate that deep learning models can effectively calibrate to and predict individual policy reserve values to provide more simplified and efficient valuations. Furthermore, the models are robust with regard to model input and can handle unseen data within valid ranges extremely well.

The results have important implications on the insurance industry. Deep learning can significantly enhance the valuation of commercial books of life insurance business. The ability to quickly and accurately value large policy portfolios can help insurance companies make better decisions with pricing, risk management and investment strategies and the automation thereof.

I provide examples where the model performs perfectly on random unseen data over many dimensions. On all model points the accuracy on individual model points are 99.5% or higher. This was obtained through introducing novel ideas at many steps throughout the process.

The results of this study also have important implications for the statistical and machine learning community, not only by the new combinations of methods developed to solve these regression problems but also as a guideline of principles that should be followed throughout the training process by practitioners.

141

Unravelling the dynamics of GGE biplots as visualisation tool to interpret and present agricultural trials

Dr Mardé Booyse[1], Dr Zelda Bijzet[1]
[1]Agricultural Research Council, Stellenbosch, South Africa

One of the challenges facing statisticians is to provide tools to enable researchers to interpret and present data and draw subsequent conclusions in ways easily understood by the scientific community for whom boxplots, histograms, and pie charts are familiar forms of data visualisation. Visualisation techniques are also more insightful and immediate as opposed to a numerical description. However, the mentioned visualisation tools available to scientists are only applicable to univariate data and therefore a graphical display for fully understanding large data sets with complex interconnectedness and interactions is required. Based on matrix mathematics, a biplot analysis can be used to interpret these complex interactions. The GGE biplot derived from the Genotype (G) plus Genotype by Environment (GxE) was developed to visualise interactions in plant breeding where multi- environment trials are prevalent. This paper will thus review the dynamics and applicability of the GGE biplots in agricultural trials using the GGE biplot visualising ability of the which-won-where pattern, the interrelationship among environments and the ranking of genotypes based on mean performance and stability. This paper will illustrate how researchers can benefit from biplots to interpret interactions in their data and then present it to their non-statistical minded audiences.

108

Multivariate stratified sampling allocation

Ms Georgi Borros[1], Dr Sebnem Er[1], Mr Sulaiman Salau[1]
[1]University of Cape Town, Cape Town, South Africa

Multivariate stratification simultaneously partitions a heterogeneous population into more homogeneous subgroups based on multiple variables of interest. Once the subgroups are formed, the sample is allocated across strata, considering multiple outcome measures simultaneously. Multivariate stratified sampling therefore involves two optimisation problems: strata boundary determination and sample size allocation across strata. This study focuses on the allocation problem in the multivariate stratification case, relevant to survey research with predetermined strata and multiple outcomes of interest. In such a case, an optimal solution for multivariate allocation is not established - as an allocation that optimises for one variable may not be optimal for the others, often deemed a 'compromise' allocation. A key characteristic of compromise allocations is an arbitrary or 'compromise' assignment of weights across outcome variables of interest. This study aims to contribute a more systematic weighting procedure using principal component analysis. Additionally, the study builds on an established random search algorithm with modifications to enhance efficiency. The updated algorithm is shown to optimise for those variables with the greatest contribution to overall variation in the dataset, thereby producing more efficient estimates relative to previous methods. The resulting solution removes the arbitrary choice of importance weights while following an intuitive optimisation search procedure.

152

New classes of tests for the Weibull distribution in the presence of random right censoring

Dr Elzanie Bothma[1], Prof James Samuel Allison[1], Prof Izak Jacobus Hennig Visagie[1]
[1]North-West University, Potchefstroom, South Africa

We develop two new classes of tests for the Weibull distribution based on Stein's method. The proposed tests are applied in the presence of random right censoring. We investigate the finite sample performance of the new tests using a comprehensive Monte Carlo study. In both the absence and presence of censoring, it is found that the newly proposed classes of tests outperform competing tests against the majority of the distributions considered. In the cases where censoring is present, we consider various censoring distributions. Some remarks on the asymptotic properties of the proposed tests are included. We present another result of independent interest; a test initially proposed for use with full samples is amended to allow for testing for the Weibull distribution in the presence of censoring. The techniques developed in the paper are illustrated using two practical examples.

155

Network analysis and disruption simulation of a South African cash-in-transit criminal network

Ms Annie Kok[1], Mr Stefan Britz[1]
[1]University of Cape Town, Cape Town, South Africa

Cash-in-transit robberies are violent crimes that continue to make headlines in South Africa for the sheer spectacle and audacity of the heists, especially vehicle-on-road robberies. The impact of these organised criminal activities is far-reaching, compromising the integrity of the cash management system, the criminal justice system, and the safety of South African citizens. Despite recent academic interest, no empirical research currently exists on the networks that organise and perpetrate these heists, partly due to the difficulty of obtaining reliable data on inherently covert operations. This research uses network analysis to describe the characteristics of a South African cash-in-transit criminal network by analysing the acquaintance testimonials, phonebook contacts, and call records as extracted from the court judgments of the notorious KZN26 case. Network-level analyses indicate that the network was not constrained by its established structure (according to degree measures), it relied on cohesive connections to maintain operational efficiency (transitivity), and that it prioritised efficiency during the robbery phase (centralisation score). Community-level analyses show that cohesion within subgroups drawn on geographic lines is stronger than between them. The node-level results also suggest that there is notable variation in the level of connection and embeddedness, which is consistent with core-periphery literature, as well as cash-in-transit crime literature in general. This shows that the manner of connections at the network level yields a well-connected core with information trickling to the peripheries, and this is shaped by their structural positions and importance. Finally, simulated disruption strategies informed by centrality measures prove to be considerably more effective at dismantling the network than random targeting.

168

bipl5: An R package for reactive calibrated axes biplots

Mr Ruan Buys[1], Ms Delia Sandilands[1], Prof Sugnet Lubbe[1], Dr Carel Johannes van der Merwe[1]
[1]Stellenbosch University, Stellenbosch, South Africa

Principal component analysis biplots with calibrated axes are popular and effective multivariate data visualisation tools. Biplots are however often complex to navigate due to cluttered plotting in the central data area, as well as limitations that accompany static rendering. The bipl5 package proposes three contributions to the biplot display: i) automated orthogonal parallel translation of the axes to the boundary of the plot and declutter the plot center; ii) superimpose interclass kernel densities on each axis to investigate class distributions in the data; iii) render the final plot on a portable and standalone HTML file with embedded reactivity. The bipl5 package is also extended to serve as a plotting wrapper for the biplotEZ package to accommodate various other biplot visualisations.

64

A consolidated approach to linear mixed models with factors having both fixed and random levels

Dr Lyson Chaka[1], Prof Peter M. Njuho[2], Prof Hans-Peter Piepho[3]
[1]University of Pretoria, Pretoria, South Africa, [2]University of Zululand, Emphangeni, South Africa,
[3]University of Hohenheim, Stuttgart, Europe

The consistent development of some crucial innovations in technology has impacted the field of statistics and data analytics, leading to the development of new strategies for handling complex experimental designs. Recent technological advancement in fields such as agriculture and other industrial processes necessitate the development of statistical modelling approaches for experiments that strive to compare the efficiency of new machinery and/or strategies against the traditional ones to establish any deviation in location or variation in the output variable. The setting dictates a linear mixed model scenario where the predictor variables (factors) are conceptualised as each made up of both fixed and random levels. Assuming a linear mixed model, the concept requires a careful consideration in model selection, parameter estimation and the assumed variance-covariance structure. The fundamental consideration in the linear mixed model framework is that, the response variable is predicted by factors whose levels are either fully fixed or random in nature. In order to obtain an improved level of precision, the proposed approach involves the partitioning of factor levels for concentrated analyses of variance based on the objectives of the experiment. Combining these partitioned analyses of variance is crucial for a broader perspective. However, the approach poses some challenges on re-arrangement of the data and coding prior to combined analyses where the current statistical software has no direct provision to handle such complexity. In addition, SAS and CRAN R environments are proposed and used to obtain the consolidated analyses.

97

On precedence tests with double sampling

Dr Niladri Chakraborty[1], Prof Narayanaswamy Balakrishnan[2], Prof Maxim Finkelstein[1]
[1]University of the Free State, Bloemfontein, South Africa, [2]McMaster University, Hamilton, Canada

We introduce and analyse double sampling-based precedence and weighted precedence tests. Within the double sampling framework, we derive the joint distributions of the two statistics. We then obtain closed-form expressions for the rejection probabilities under both the null hypothesis and the Lehmann alternative. A power comparison is conducted against the Lehmann and location-scale alternatives using Monte-Carlo simulations.

137

Modelling divest South African stock prices using mixture distribution

Prof Martin Chanza[1,2,4], Dr Modisane Seitshiro[3,4]
[1]Technology Enhanced Learning and Innovative Education and Training in South Africa (TELIT-SA), North-West University, Vanderbijlpark, South Africa, [2]Department of Statistics and Operations Research, North-West University, Mafikeng, South Africa, [3]Centre for BMI, North-West University, Potchefstroom, South Africa, [4]National Institute for Theoretical and Computational Sciences (NITheCS), South Africa

Skewness and kurtosis for stock price analysis pose uncertainty. The stock price distribution must be clearly defined and understood to address these uncertainties. This study aims to fit different probability distributions conforming to the stock prices. It uses daily stock price data from January 2015 to June 2024 to model the probability distributions of stock returns for three well-known South African companies. Mixture distribution, Normal, Student's t, gamma, and Pareto distributions are used to assess the uncertainty of the stock returns. The results shed light on the uncertainty of stock returns, such as stock price volatility and asymmetric returns behaviour. This analysis offers valuable insights into selecting appropriate distributional frameworks for accurately modelling stock price movements in emerging markets. This will shed light on selecting appropriate distributions for fund managers, traders, and risk managers.

40

Experimental designs for estimating non-linear models in mixture variables

Prof Roelof Coetzer[1]
[1]North-West University, Potchefstroom, South Africa

Mixture variables are common in various applications such as food sciences, ecology, agriculture, chemistry, fuel blending, and feed and product compositions in chemical processes. Scheffè S- and K-polynomials are well known for modelling the properties of mixtures. However, in some applications such as in studies of physical properties of liquids and pressure drop in chemical reactors, non-linear models must be specified and employed for estimating the responses as a function of the mixture variables. In this paper, non-linear mixture models such as weighted power-mean mixture models and Padè approximations, which are ratios of K-polynomials, will be presented and discussed. Specifically, mixture designs and D- and I-optimal designs will be discussed for estimating Padè polynomials.

69

Enhanced process monitoring: the bootstrapped cumulative sum-exponentially weighted moving average (BCUSUM-EWMA) control chart approach

Dr Braimah Joseph Odunayo[1], Prof Fabio Correa[1]
[1]University of the Free State, Bloemfontein, South Africa

This study addresses challenges in Phase II univariate process control, where in-control data exists but underlying process distributions are unknown. Traditional control charts often require specific knowledge of these distributions, which is impractical in many real-world applications. This paper proposes novel control charts, the Bootstrap Cumulative Sum-Exponentially Weighted Moving Average (BCUSUM-EWMA) charts, designed for any process (mean and standard deviation) monitoring. These charts utilize bootstrapping to overcome limitations imposed by normality assumptions, which may not hold true in practice. For comparison, bootstrap versions of CUSUM (BCUSUM) and EWMA (BEWMA) charts are also developed. The performance of these charts is evaluated using Average Run Lengths (ARLs) calculated via Monte Carlo simulation in R software. To demonstrate the practical application of our proposed BCUSUM-EWMA control chart, we analysed real-world wearer heart rate data from 37 patients collected from the record office at Irrua Specialist Teaching Hospital. We employed a bootstrap simulation of 1500 samples to evaluate the chart's performance. The proposed methods are then demonstrated with real-world wearer heart rate data. Compared to classical control charts, the bootstrap charts signal out-of-control shifts earlier. Additionally, performance assessment based on ARLs confirms the effectiveness of the bootstrap approach, with smaller out-of-control ARLs indicating earlier detection.

101

Seasonal catchment areas using an attribute based fuzzy lattice data structure

Mrs Michelle de Klerk[1], Prof Inger Fabris-Rotelli[1]
[1]University Of Pretoria, Pretoria, South Africa

Seasonality impacts various industries and sectors, influencing agricultural cycles, economic planning, and healthcare resource allocation. We propose a novel approach using an attribute based fuzzy lattice data structure to create overlapping catchment areas using the fundamentals of label propagation and graph clustering. This approach considers both the link structure and attribute similarities between nodes in a network, where the nodes are points of interest in a road network. Nodes may be close or far apart based on connectivity and shared attributes, such as common interests or in a geographical application considering topography features. In this study, we incorporate static and seasonal attributes for geographical nodes, allowing us to explore seasonal catchment areas and provide a more realistic view of accessibility throughout the year. This integrated approach offers a comprehensive framework for assessing spatial accessibility and understanding seasonal variations in regions to enhance planning for essential services.

167

Shewhart X control charts for monitoring the mean of autocorrelated AR(1) data

Dr MD Diko[1], Mr Tiisetso Molele[1], Mr Ntlhari Mabunda[1], Mr Matsebe Mabotha[1]
[1]University of the Free State, Bloemfontein, South Africa

Traditional Shewhart control charts typically assume independent data, but in reality, data is rarely independent. Applying a traditional Shewhart chart to autocorrelated data can increase the false alarm rate. This study examines the impact of autocorrelation on the performance of traditional individual Shewhart charts. To address the effects of autocorrelation, several alternative Shewhart control charts have been proposed, including the residual Shewhart chart, the modified Shewhart chart, and the modified residual Shewhart chart. These charts are compared for various AR(1) processes by simulating run-length distributions and average run-length curves using the software R. The results indicate that for low to moderate positive autocorrelation, the modified control chart is most effective, while for very high positive autocorrelation, the modified residual Shewhart chart performs best. For negative autocorrelation, the residual Shewhart chart is preferred.

136

Design of spatial capture recapture (SCR) surveys for stratified populations

Dr Greg Distiller[1,2], Associate Prof Ian Durbach[1,2], Ms Anita Wilkinson[3]
[1]Department of Statistical Sciences, University of Cape Town, Cape Town, South Africa, [2]Centre for Statistics in Ecology, the Environment, and Conservation, Cape Town, South Africa, [3]The Cape Leopard Trust, Cape Town, South Africa

Spatial Capture-Recapture (SCR) models use data collected by an array of detectors to model animal density. Despite the considerable time and effort that setting up the detectors typically takes, there has been limited guidance on the best way to configure the array. However recent research has explored an approach whereby different criteria are optimised to generate proposed designs. This work generates designs for a single population, yet many species exhibit sex-specific differences. We explore how the existing algorithms perform on a stratified population and evaluate the effect of using information from both strata when generating designs. Our case study is based on data from a real survey conducted by the Cape Leopard Trust in the Boland Mountain complex."

50

Understanding the impact of parameter estimates on model performance

Lindani Dube[1,3], Tatenda Shoko[2], Tanja Verster[1,3]
[1]North-West University Centre for BMI, Potchefstroom, South Africa, [2]AIMS, Stellenbosch, South Africa, [3]NITheCS, Stellenbosch, South Africa

This research will explore how Explainable Machine Learning (XML) can be used in credit risk management. It will focus on Partial Dependence Plots (PDPs), which help visualise the relationship between input variables and credit risk predictions. PDPs are easy to understand and work with any model, showing how different factors affect the outcomes. This study will compare the use of PDPs with traditional methods like logistic regression to see how they improve model transparency and performance. By applying PDPs to various machine learning models, this research aims to enhance transparency, meet regulatory requirements, and build trust with stakeholders in the financial industry.

164

Exploratory spatial analysis of early grade reading data in KwaZulu-Natal

Mr Joshua Engelbrecht[1]
[1]JET Education Services, Johannesburg, South Africa

This research uses spatial econometrics to explore the geographic patterns related to early grade reading outcomes across five districts in KwaZulu-Natal, using data from the Early Grade Reading Intervention (EGRI) pilot. The research analyses data from 5 192 learners and 162 schools participating in the pilot, investigating spatial relationships between schools and across districts, in order to investigate their impact on literacy outcomes.

Spatial econometrics methods account for geographic relationships and dependencies in data analysis. The purpose of this research is to experiment with spatial techniques to determine if they can form a foundation for predictive modelling. This is motivated by the potential to accurately predict learner literacy levels in low-income schools while balancing data requirements, costs, and availability, so that the targeting of policy and interventions related to literacy can be improved.

The presence of spatial relationships is investigated using Moran's I for spatial autocorrelation, revealing a Moran I statistic of 8.45 ($p = 0.0022$), indicating significant spatial clustering of literacy scores. Literacy scores are modelled using Spatial Error (SEM), Spatial Lag (SLM), and Spatial Durbin (SDM) models, with the independent variable being quintile, an ordered indicator of school SES.

Results show that early grade reading outcomes exhibit significant spatial dependence, with high-performing clusters identified in two districts and 80/162 schools categorised as insignificant across the remaining three districts. The SEM model indicates a positive association between quintile and literacy outcomes (Estimate: 2.16, $p = 0.039$), suggesting that higher SES correlates with improved reading performance. The SDM model further demonstrates that neighbouring schools' quintile influences performance (Estimate: -18.25, $p = 0.046$), reflecting spatial interaction dynamics. Overall, the findings underscore the significance of spatial effects on early grade reading outcomes, highlighting the importance of incorporating geographic data to enhance statistical modelling for literacy interventions.

142

A methodology for wave detection in epidemics

Dr Warren Brettenny[4], Nada Abdelatif[1], Jenny Holloway[2], Nontembeko Dudeni-Tlhone[2], Prof Inger Fabris-Rotelli[3], Prof Pravesh Debba[2], Dr Sibu Makhanya[5], Dr Raeesa Docrat[6], Wouter le Roux[2]
[1]South African Medical Research Council, [2]Council for Scientific and Industrial Research, [3]University of Pretoria, [4]Nelson Mandela University, [5]IBM South Africa, [6]University of the Witwatersrand

In both the management and modelling of epidemic outbreaks, the ability to determine the start of a wave of cases is of vital importance. Not only does this advantage the modelling of the outbreak, but if done in real time, can assist with a nation's response to the disease. In this study, a bidirectional long-short-term-memory (BD-LSTM) network is used to determine the start and end of the COVID-19 waves experienced in the districts and metros of Gauteng, South Africa, from 2020-2022 as well as the waves of the cholera outbreaks occurring in the Beira area of Mozambique between 1999 and 2005, in real-time. Using these starting predicted dates, effective spatial-SEIR models are used to predict the trajectory of the COVID-19 cases in each district and metro of Gauteng. The fitted BD-LSTM model demonstrates that it is effective in predicting wave start and end dates in real-time and its use is not limited to COVID-19 studies but can also be applied to other disease outbreaks.

21

Trends in quantity and demographic composition of statistics graduates at South African universities, 1986-2022

Dr Thomas Farrar[1]
[1]Cape Peninsula University of Technology, Bellville, South Africa

Publicly available enrolments and graduates headcount data from the Department of Higher Education and Training website was accessed, aggregated, and transformed to create a single dataset spanning the period 1986 to 2022. Descriptive analyses were conducted of enrolments and graduates in probability and statistics using CESM categories. Explanatory variables available in the data include gender, population group, qualification type, NQF level, and institution. The data indicate long-term growth in probability and statistics graduates in absolute numbers. They also indicate long-term transformation in the demographic composition of these graduates toward greater representativity of the South African population. However, some demographic groups remain underrepresented in enrolments and graduates. Furthermore, demographic transformation has been slower at higher levels of study and has also levelled off in recent years. Proxy measures for throughput were obtained from the data and these indicate that historically disadvantaged demographic groups have lower throughput. Thus, despite progress made in the past three decades, the discipline of statistics in South African higher education is not yet close to achieving equity in terms of access or outcomes.

46

Development and implementation of fictional narratives for enriched teaching of university-level statistics

Prof Johan Ferreira[1], Dr Seite Makgai[1]
[1]University of Pretoria, Pretoria, South Africa

In a world where problems become more complex to solve, statistics remain a scary and uninviting field of science, even beyond introductory courses. Few existing teaching approaches focus on investigating the potential and impact of facilitating deep-learning of concepts covered in advanced undergraduate courses in a drastically creative way. We recognised a potential void in the South African tertiary statistics curriculum and set out to supplement the traditional theoretical (and practical) praxis with a fresh perspective. Time series analysis is an 18 credit third year module that is taken by approximately 300 students annually at the University of Pretoria, consisting of a student body originating from diverse backgrounds within different faculties. We develop and curate a collection of fictional narratives (short stories) that are based on content from this syllabus. Students voluntarily participated in a storytelling exercise, where key time series analysis concepts were characters, to explore the stimulation of previously unconsidered cognitive centres that might supplement another such centre - to accelerate learning and decrease perceived anxiety surrounding the study of time series analysis. The outcome of this study is threefold: first, we review literature from pedagogy pertaining to creative thinking within analytical sciences at the tertiary level; motivating the exploration of fictional narratives in advanced undergraduate statistics courses, shedding light on alternative and less stressful teaching and learning approaches for students. Secondly, an open-access contributed volume of student-penned fictional narratives is published, which may serve as an additional learning resource within an introductory time series analysis syllabus. Finally, an initial qualitative-type analysis investigates students' experience of this potential additional learning resource. Initial results indicate that formal learning is facilitated in this informal and previously unconsidered peer learning way that supports creative, project-based, and transdisciplinary learning.

62

Mapping linguistic beauty: biplot analysis of 228 world language patterns

Dr Raeesa Ganey[1,2], Dr Johané Nienkemper-Swanepoel[2,3]
[1]School of Statistics and Actuarial Science, University of Witwatersrand, Johannesburg, South Africa, [2]Centre for Multi-Dimensional Data Visualisation (MuViSU), Stellenbosch University, Stellenbosch, South Africa, [3]Department of Statistics and Actuarial Science, Stellenbosch University, Stellenbosch, South Africa

In early 2024, The Economist released an article, "What is the world's loveliest language?", which delves into the subjective perceptions of linguistic beauty. This was based on a publication from Anikin, Aseyev & Johansson (2023). The piece discusses how various cultures and individuals view languages differently. These perceptions are influenced by a mixture of phonaesthetics qualities, which are emotional responses to sounds and acoustic properties like vowel harmony, consonant softness, and rhythmic flow. The results underscore that language beauty is subjective, shaped by both cultural familiarity and inherent linguistic traits. This presentation utilises multivariate data from 228 languages by analysing phonaesthetics and acoustic measurements to objectively assess aesthetic judgements from the original publication. The aim of this presentation is to show how a variety of multivariate visualisations, specifically biplots, enhances the interpretations of the results and exposes complex interactions between linguistic features and aesthetic perceptions.

Reference

Anikin, A., Aseyev, N. & Johansson, N.E. 2023. Do some languages sound more beautiful than others? Psychological And Cognitive Sciences (PNAS). 120(17): 1-7. https://doi.org/10.1073/pnas.2218367120

130

Type I multivariate Pólya-Aeppli distributions with applications

Ms Claire Geldenhuys[1], Dr René Ehlers[1], Prof Andriette Bekker[1]
[1]Department of Statistics, University Of Pretoria, South Africa

An extensive body of literature exists that specifically addresses the univariate case of zero-inflated count models. In contrast, research pertaining to multivariate models is notably less developed. We proposed two new parsimonious multivariate models which can be used to model correlated multivariate overdispersed count data. Furthermore, for different parameter settings and sample sizes, various simulations are performed. In conclusion, we demonstrated the performance of the newly proposed multivariate candidates on two benchmark datasets, which surpasses that of several alternative approaches.

57

Dynamic prediction and standard prediction models for type 2 diabetic individuals in the Western Cape

Mr Frissiano Honwana[1], Associate Prof F Gumedze[1], Prof L Myer[1], Dr J Rusch[2]
[1]University of Cape Town, Cape Town, South Africa, [2]National Health Laboratory Service, Cape Town, South Africa

Background
Dynamic risk prediction models can be a key tool for personalized medicine. Joint modelling approaches are often used for individualized dynamic risk predictions and are typically regarded as a superior approach compared to Cox-based risk predictions. However, this superiority is not always true across all contexts. This study compared the predictive performance of joint modelling and Cox-based prediction models for glycaemic control in individuals with type 2 diabetes (T2DM) in the Western Cape, South Africa.

Methods
Routine data from 105 011 T2DM individuals in public healthcare facilities across the Western Cape from 2016 to 2021 were analysed. The data was split into a development cohort (n = 78 868) and a validation cohort (n = 26 143). Joint modelling and Cox-based modelling were compared in terms of discrimination and calibration using appropriately formulated repeated haemoglobin A1c (HbA1c) measurements. The models' discrimination and calibration were compared using the area under the curve (AUC) of the receiver operating characteristic (ROC) and Brier scores, respectively.

Results
The Cox-based prediction model demonstrated comparable predictive performance to the joint modelling approach, with AUCs of 0.65 vs. 0.67 and Brier scores of 0.065 vs. 0.074, respectively.

Conclusion
While joint modelling, which incorporates repeated biomarker measurements, has potential for individualized prognosis prediction, it does not always outperform the Cox-based approach, as seen in the routine data from a resource-limited setting. These findings underscore the need for understanding factors influencing predictive performance across varying contexts and data characteristics for risk predictions.

150

Identifying differences between batters in Twenty20 cricket using principal component analysis and biplots

Mr Cameron Howe-Dreyer[1], Dr Paul J van Staden[1], Prof Inger N Fabris-Rotelli[1]
[1]Department of Statistics, University of Pretoria, Pretoria, South Africa

This study analyses the batting performance of cricketers in Twenty20 cricket. In order to do so, a more comprehensive statistical analysis is proposed instead of the basic statistics at measuring players typically used in cricket, such as batting averages and strike rates. The research proposes applying principal component analysis (PCA) to certain variables describing a batter. The aim of PCA is to reduce the dimensionality of the data, and the resulting principal components are then graphically represented using biplots. A biplot visually shows the relationships between variables and observations which can reveal any clusters or multicollinearity to guide the analysis. For this study, ball-by-ball data on the Indian Premier League (IPL) of years 2018-2023 is extracted. The results from different batters on the biplots show strong clustering of observations for the three stages of a Twenty20 innings, namely the first powerplay (overs 1-6), the middle overs (overs 7-14) and the final overs (overs 15-20), indicating that batting orders in Twenty20 cricket are important.

170

Seasonal volatility patterns in SAFEX grain futures: analysing environmental and supply-side risks

Dr Ayesha Sayed[1], Associate Prof Chun-Sung Huang[1]
[1]Department of Finance and Tax, University of Cape Town, Cape Town, South Africa

Climate change is reshaping agricultural production and market dynamics, particularly for crops like maize, which are vital to global food security. This paper examines how shifts in the planting, growing, harvesting, and marketing stages of maize cultivation influence volatility in futures contracts. In the Southern Hemisphere, maize is planted from October to December, grows through the summer months, and is harvested from March to May. These stages are increasingly disrupted by climate change, with implications for futures markets. Unpredictable weather patterns, such as altered precipitation and temperature extremes, lead to variations in planting and harvesting schedules, affecting supply estimates and increasing market uncertainty. Delays or advancements in planting, unexpected changes in crop growth rates, and quality issues at harvest can all contribute to volatility in futures prices. Additionally, post-harvest factors, including supply chain disruptions and storage challenges, further exacerbate market fluctuations. This paper employs a comprehensive analysis of historical market data, climate impact studies, and futures trading patterns to illustrate how climate-induced disruptions propagate through the maize supply chain and affect futures contracts. By highlighting the interplay between climate variability and market volatility, this research aims to provide insights into managing risks associated with maize futures and offers recommendations for traders, policymakers, and stakeholders in adapting to an evolving agricultural landscape.

126

Stakeholder focused explainable artificial intelligence

Ms Gandhi Jafta[1], Prof Inger Fabris-Rotelli[1], Prof Emma Ruttkamp-Bloem[1], Dr Iketle Maharela[1]
[1]University of Pretoria, Pretoria, South Africa

Integrating Artificial Intelligence (AI) in high-consequence human environments has raised concerns about the fairness, accountability, transparency, and decision-making processes of AI models. Explainable Artificial Intelligence (XAI) at its inception was conceptualised as a suite of machine learning techniques that can explain the decisions made by machine learning algorithms. However, existing XAI approaches have limitations in addressing users' understanding and fail to consider the perspectives and requirements of different stakeholders. This research proposes a novel stakeholder-centred XAI framework that incorporates insights from social sciences to provide understandable and satisfactory explanations to stakeholders in high-consequence environments. The novel solutions proposed in this research are incorporated into each of its three parts:

1. Exploring the significance of XAI and examining how it can address fairness, and bias concerns for all stakeholders involved. The stakeholders include end users, regulators, governments, domain experts, decision-makers, and decision recipients.
2. Formulating and formalising different types of explanations needed by the different stakeholders, with a particular focus on examining machine learning techniques used in high-consequence environments.
3. Emphasising the importance of AI and data literacy for all stakeholders. AI and data literacy should be considered a public good—a resource that is both non-excludable and non-rivalrous. Once knowledge and educational resources are made accessible, they can benefit society as a whole without being depleted, thereby promoting informed and equitable participation in an increasingly data-driven world.

This research contributes to developing a user-centric XAI framework, bridging the gap between technical expertise and social understanding, and promoting trust, transparency, and accountability in AI systems, and fostering multi-disciplinary engagement in XAI. Moreover, it significantly contributes to machine learning and statistics by proposing novel XAI techniques and methodologies. The outcomes of this research will foster the responsible and trustworthy use of AI, positively impacting high-consequence human environments and facilitating informed decision-making.

93

Modelling toroidal data for representation and analysis of protein dihedral angles

Mr Claudio Jardim[1], Prof Inger Fabris-Rotelli[1], Dr Alta de Waal[1], Dr Najmeh Nakhaei Rad[1]
[1]University Of Pretoria, Pretoria, South Africa

We introduce a comprehensive approach to modelling bivariate toroidal data by applying more flexible distributions that account for the circular nature of angular data. Dihedral angles, pivotal in defining protein conformation, are one of many occurrences of toroidal data in scientific fields. By adopting these improved models, we can achieve a more precise and nuanced representation of the structural representation of proteins. This refined structural representation enhances the accuracy of computational methods, from structure prediction to understanding the dynamic behaviour of proteins. This improvement in computational methods and structural understanding can lead to advancements in protein engineering and improve drug design. We perform an in-depth analysis by employing specific probability distributions and estimating parameters. We investigate and compare these methodologies using protein dihedral angle data. This allows us to pinpoint the most suitable model for each amino acid residue, resulting in a comprehensive understanding of the effectiveness of our techniques.

82

The impact of environmental shocks due to climate change on intimate partner violence: a SEM

Ms Esme Jordaan[1,2], Ms Jenevieve Mannell[3,4]
[1]Biostatistics Research Unit, South African Medical Research Council, Parow, South Africa, [2]Statistics and Population Studies, University of the Western Cape, Cape Town, South Africa, [3]Institute for Global Health, UCL, London, UK, [4]National University of Samoa, Samoa

The impact of climate change on human societies is now well recognised. However, little is known about how climate change alters health conditions over time. National level data around climate shocks and subsequent rates of intimate partner violence (IPV) could have relevance for resilience policy and programming. We hypothesise that climate shocks are associated with a higher national prevalence of IPV two years following a shock, and that this relationship persists for countries with different levels of economic development. We also investigate potential indirect effects on IPV. We compiled national data for the prevalence of IPV from 363 nationally representative surveys from 1993 to 2019. These representative data from ever-partnered women defined IPV incidence as any past-year act of physical and/or sexual violence. We also compiled data from the Emergency Events Database on the national frequency of eight climate shocks from 1920 to 2022 within 190 countries. Using exploratory factor analysis, we investigated the number of climate shock factors and loadings of the climate shock variables using fit indices to find the best model. To assess the hypothesis, a structural model with IPV regressed on the resulting climate shock factors was done. To investigate the effect of economic development, we included GDP to the structural model. In a subsequent model we added indirect effects on IPV, mediated by gender inequality and total alcohol consumption per capita. National data representing 156 countries suggest a significant relationship between IPV and a climate factor (Hydro-meteorological) composed of storms, landslides and floods (standardised estimate=0·32; SE=0·140; p=0·022). Other climate factors (Geological /Atmospheric) had no measurable association with IPV. GDP has a moderately large association with IPV (estimate=-0·529; SE=0·065; p=0·0001). Model fit overall was satisfactory (RMSEA=0·064 (90%CI: 0·04-0·08); CFI=0·91; SRMR=0·063). Significant indirect effects will be reported.

48

Emailed publication invitations received by biostatisticians: academically sound versus potentially predatory journals

Prof Gina Joubert[1], Dr Omololu Aluko[1]
[1]University of the Free State, Bloemfontein, South Africa

Researchers increasingly receive publication invitations via email. We analysed emailed publication invitations received by staff members of the Department of Biostatistics, University of the Free State (UFS), via UFS email accounts from May to July 2023, comparing emails from accredited and non-accredited journals. Two researchers independently completed the data form for each email, then checked and resolved any discrepancies. Three of the six staff members received a total of 129 publication invitations during the study period, of which the majority (86.8%) were regarding specific journals. Ninety-three distinct emails were received from 88 journals. Two of these journals related to biostatistics. Fifteen emails (16%) were received from a journal appearing on the Department of Higher Education and Training (DHET) accredited journal lists. Emails from non-accredited journals were significantly ($p<0.01$) less likely to: refer to a journal with a health sciences-related title (37% versus 86%), indicate the publisher (36% versus 93%), provide a link to the journal website (59% versus 100%), state a full physical address (24% versus 80%), refer to author instructions (21% versus 47%) or request the recipient to share the email with colleagues (5% versus 47%). Emails from non-accredited journals were significantly ($p < 0.01$) more likely to: contain grammatical errors (63% versus 0%) and flattering remarks regarding the recipient or his/her research work (49% versus 0%), and to indicate the journal's ISSN number (67% versus 13%). In conclusion, publication invitations by email were received from accredited journals, not all such invitations are thus suspect. The clear differences between email invitations from accredited versus non-accredited journals provide insight into warning signals.

106

A stochastic modelling of South African COVID-19 mortality, new infections and vaccination dynamics

Mr Malandala Kajingulu[1], Prof Edmore Ranganai[1]
[1]University of South Africa, Johannesburg, South Africa

Modelling COVID-19 mortality is crucial for shaping effective public health strategies. This study uses COVID-19 data from March 2020 to May 2022, applying the Generalized Additive Model for Location, Scale, and Shape (GAMLSS), an advanced extension of Generalized Linear Models (GLM) and Generalized Additive Models (GAM). GAMLSS explores the dynamics of daily mortality, new infections, and vaccination rates in South Africa. Its advanced smoothing capabilities allow for a more flexible and precise analysis of count data. The study uncovers a significant inverse relationship between vaccination rates and mortality. These findings provide a strong foundation for optimizing public health policies and vaccination strategies to better manage COVID-19 transmission and reduce fatalities.

27

Non-parametric methods for forecasting South African maize and wheat prices

Ms. Emelia Kammies[1]
[1]Sol Plaatje University, Kimberley, Northern Cape, South Africa

Price fluctuations affect livelihoods and social stability, thus it is crucial for farmers, distributors, processors, and policymakers to study price behaviour and predict future trends when planning agricultural activities. Over time, various forecasting methodologies have been developed to predict grain prices. However, some of these methodologies such as the ARIMA and neural network do not produce accurate forecasts due to some inherent factors. To address these issues, this study aims to use non-parametric models, like Dynamic Generalised Additive Models (DGAMs), to predict future grain prices in South Africa. Historical time series data from RSA SAFEX Domestic future prices spanning the period 1996 to 2024 and factors such as weather variables, exchange rate, and fuel prices from appropriate sources are used in fitting the DGAMs model. The results will then be compared to traditional ARIMA models to assess the performance of the DGAMs model. Dynamic Generalised Additive Models are more flexible thus they can capture complex patterns and may yield accurate forecasts. The results of the study are intended to provide farmers and stakeholders with price certainty and the impact of factors affecting the price for planning and policy directions on tackling food insecurity.

10

Virtual screening of plants and compounds against various disease targets using machine learning

Mr Alexander Kelbrick[1], Dr Najmeh Nakhaei Rad[1], Dr Paul van Staden[1], Prof Vinesh Maharaj[2]
[1]University of Pretoria, Pretoria, South Africa, [2]Department of Chemistry, University of Pretoria

Many drugs used today are inspired by nature. Discovering active compounds within plants that target specific diseases paves the way for developing new medicines. This research outlines the use of word embedding-based virtual screening (WEBVS) to achieve two main aims: predicting novel useful compounds from plants, and also predicting the viability of known compounds for beta-lactamase inhibition. The main goal of WEBVS is to reduce the time it takes scientists to discover new useful plants or compounds. Few computational approaches for finding viable plant extracts have been developed in the literature, with Bio-assay being the primary method employed. However, Bio-assay is a time-consuming and often manual process. WEBVS aims to supplement this approach by using available textual data from literature. In this study, the abstracts of over 400000 papers were used to generate word embeddings for two problems. A continuous bag-of-words (CBOW) architecture was used to encode the literature. For compounds, we used data annotated on the Chemical Entities of Biological Interest (ChEBI) database. Labelled plant data was retrieved from the Collective Molecular Activities of Useful Plants database, plants labelled as anti-malarial were considered active. For beta-lactamase inhibition, a compound was regarded as active if it appeared in a search for beta-lactamase inhibitors. The anti-malarial model had a precision of 43.62% with a recall of 92.56% for the top 25 results. Our anti-microbial model (beta-lactamase inhibition) achieved a precision of 96.67% with recall of 89.12%. These results were based on 100 bootstrap samples of training and test sets. A shortlist of the predicted plants and compounds was assessed using Bio-assay.

71

A discrete-time competing risk analysis of students' academic behaviour: cause-specific and subdistribution hazards approach

Dr Lionel Establet Kemda[1]
[1] Durban University of Technology, Durban, South Africa

Student academic performance is an essential part of higher learning institutions. Predicting student academic performance becomes more challenging due to the large volume of data in educational databases. However, academic performance is hindered by high dropout rates. This paper builds cause-specific odds and subdistribution hazard models that describe a students' progression in a university from first registration until dropout or degree completion occurs. Based on that, school administrators can take the necessary steps to improve students' academic performance. Students' academic data involving first-time entrants into the College of Agriculture, Engineering and Science at the University of KwaZulu-Natal, South Africa, is analysed using the discrete-time competing risks specifications. The data include male and female students with over 80% aged between 17 and 19 years. According to our findings, the number of matriculation points, having financial aid, and semester weighted average contribute positively to the cause-specific and the subdistribution hazards of degree completion. However, being in a university-type of residence and students' gender only affects the odds of dropping out, but not the odds of degree completion. It was also established that the number of matriculation points obtained rather than the type of quintile school attended had a statistically significant effect on the cause-specific odds of degree completion. Additionally, students from science had a higher dropout rate, so the university can learn about the students' behaviour of different faculties and provide corresponding personalized services, which have particular practical significance.

22

Taking data science collaboration to new heights in a study to better understand perceived versus actual digital behaviour

Mr Fallo Happy Khanye[1,2], Prof Renette Blignaut[1], Dr Julia Keddie[1]
[1]University of the Western Cape, Bellville, South Africa, [2]Ghent University, Ghent, Belgium

This paper intends taking data science collaboration to the next level with multi-disciplinary, multi-institution, multi-country and multi-focus collaboration. The Digital Inclusion South Africa (DISA) questionnaire is a tool developed by the Western Cape Colab in partnership with Media, Innovation and Communication, a research group from Ghent University (imec-mict-ugent) and the Department of Statistics and Population Studies at the University of the Western Cape, which aims to measure the digital inclusivity and digital ability of South Africans. Imec-mict-ugent has an existing tool called the Digimeter that measures the digital inclusivity of Flemish people in Belgium. This presentation will report on the questionnaire development process for a South African version that will reflect a South African context. A pilot study was conducted to test the initial version of the DISA questionnaire and after several workshops and statistical analysis, some questions needed to be rephrased to reflect a South African context. Theoretical constructs were used as part of the development of the DISA questionnaire so construct validation was required using the pilot study data. We discovered that some of the constructs such as coping strategies and confidence in digital technology had low Cronbach alpha values, which indicated questionable constructs. These results were used to guide the finalisation of the national DISA questionnaire which aims to measure participants' perceived digital behaviour. The actual digital behaviour will be measured using a smartphone logging application called mobileDNA. mobileDNA will be downloaded by questionnaire participants and data collected for a two-week period. The DISA questionnaire and mobileDNA app will give rich information to compare perceived to actual digital behaviour. Our experiences, challenges and recommended resolutions related to the mobileDNA pilot study, will be discussed.

66

Entropy penalised self-paced learning

Mr Andre Ruben Kleynhans[1], Prof Frans Kanfer[1], Prof Sollie Millard[1]
[1]University of Pretoria, Pretoria, South Africa

Self-paced learning (SPL) is an algorithm that can be used to obtain a maximum likelihood estimate of a finite mixture model. SPL mitigates the impact of non-typical observations by systematically introducing observations in a meaningful order. SPL does this by considering an observation's contribution to the overall log-likelihood. We propose the use of an entropy penalty term in the SPL objection function as the regularisation term compared to a hard regularisation term for observation subset selection. We demonstrate properties of the approach using a simulation study and an application on real data.

144

Economic recession prediction using modified gradient boosting and principal component neural network algorithms

Mr Anuroop Krishnannair[1], Dr Najmeh Nakhaei Rad[1]
[1]University of Pretoria, Pretoria, South Africa

In the ever-evolving landscape of global economics, predicting and understanding economic recessions remain critical challenges for policymakers, researchers, and financial analysts. The outbreak of the COVID-19 pandemic in 2019 introduced unprecedented complexities, reshaping the economic dynamics of nations worldwide. This research presents the most effective models to assist businesses in predicting recession periods and identifies the key variables to improve the models' overall performance. To achieve this, in addition to artificial neural networks (ANN), machine learning techniques such as random forests and support vector machines are employed to develop an efficient prediction model aimed at mitigating government deficits, growing inequality, declining incomes, and rising unemployment. Furthermore, an ensemble approach combining Logistic Regression and Non-Linear Principal Component Analysis (NLPCA-LR) is proposed and compared to these models. An adjusted Gradient Boosting Neural Network (GBNN) is also proposed and compared to other models. A real dataset of historical recession periods in African countries is used to demonstrate the performance of the algorithms in practice. The findings underscore the superior performance of the Gradient Boosting Neural Network (GBNN) and Non-Linear PCA Logistic Regression Ensemble (NLPCA-LR) models across the most significant predictive metrics.

133

Bayesian approach to the estimation of asymptotic dependence and independence in joint tails

Mr Nicholas Kwaramba[1], Prof Andrehette Verster[1]
[1]University of the Free State, Bloemfontein, South Africa

In this study we consider the modelling of multivariate extreme values. We revisit the paper of Ramos and Ledford (2009) that derived a joint survivor function that considers asymptotic dependence and independence in joint tails. The model introduces three unique parameters. It also proves to work in weak or negative association cases.

We propose a Bayesian approach to the Ramos and Ledford (2009) model. Some prior information regarding the extreme value index is used to simplify the simulation process from the joint posterior. The Bayesian model is tested through simulation studies and applications to real data sets. Joint tail probabilities and posterior predictive distributions, of possible unobserved values conditional on the observed values, are also considered.

73

A combined point process for better-than-minimal, minimal, and worse-than-minimal repairs

Miss Amy Langston[1], Prof Maxim Finkelstein[2,3], Prof Ji Hwan Cha[4]
[1]Rhodes University, Makhanda, South Africa, [2]University of the Free State, Bloemfontein, South Africa, [3]University of Strathclyde, Glasgow, Scotland, [4]Ewha Womans University, Seoul, Republic of Korea

Considerable attention in reliability literature has been given to studying various repair models. In line with this, we introduce a new combined repair process to describe repairs that are initially better-than-minimal, then become minimal, before finally becoming worse-than-minimal, thus creating the corresponding three-phase bathtub pattern. The recently characterised self-regulating (with negatively dependent increments) extended generalized Polya process (EGPP), the non-homogeneous Poisson process (NHPP), and the self-exciting (with positively dependent increments) generalized Polya process (GPP) are used to describe these phases, respectively. Several useful stochastic properties of the proposed model are described, and the corresponding analytical results are derived for the combined process under two settings: change in repair type after a specified time and change in repair type after a specified number of failures/repairs. We discuss the theoretical novelty and practical usefulness of the obtained results. Specifically, as an application, the optimal age replacement problem is defined, and its optimal solution is analysed. Detailed numerical examples support our findings.

135

Logratio analysis (LRA) and compositional biplots of milk fatty acids

Dr Susan Laurens[1], Dr Raeesa Ganey[2]
[1]Heineken Beverages, Stellenbosch, South Africa, [2]University of the Witwatersrand, Johannesburg, South Africa

Compositional Data Analysis (CoDA), or logratio analysis (LRA), is a statistical approach that examines the ratio between variables after log transformation. This methodology is necessary when dealing with compositional data, which consists of vectors with non-negative values, summing to one. In such cases, traditional statistics, such as mean and variance, are not directly applicable. CoDA allows for the analysis of p variables (parts) by calculating $p(p-1)/2$ possible pairwise logratios (PLRs). It is shown that a set of $p-1$ independent logratios is sufficient to represent the full structure of the data. This study applies LRA to a compositional dataset of 74 milk fatty acids (FAs), aiming to reveal insights about their relationships.

By utilising transformations like centred logratio (CLR), PLR, and additive logratio (ALR), the dataset is visualised by compositional biplots. LRA is the PCA of all the $p(p-1)/2$ logratios and is mathematically equivalent to the PCA of the p CLRs. A compositional biplot is constructed from the p CLRs. This biplot offers an enhanced understanding of the relationships between the FAs. PCA of the $p-1$ ALRs, where the ratios are displayed as the variables, results in a biplot which is almost identical to the Compositional biplot of the full dataset, depending on the reference part. This set of $p-1$ ALRs is sufficient to describe the full structure of the data. By analysing logratios, rather than absolute values, CoDA effectively captures the relative importance of FAs. The visualisation of the milk FA data by compositional biplots offers valuable insights into their interactions and illustrates how different milk FAs contribute differently across various feeding regimes of the dairy cow.

124

Timeseries PCA biplots

Prof Sugnet Lubbe[1]
[1]Stellenbosch University, Stellenbosch, South Africa

While Principal Component Analysis (PCA) is a popular and simple method for dimension reduction, when applied to multivariate timeseries data, the temporal order of the data is not taken into account. PCA biplots are useful to represent samples and their relationship with multiple variables visually as a multidimensional scatter plot. Since PCA does not take any ordering of the observations into account, a PCA biplot of multivariate timeseries conveys no information on the temporal order of the data. In this paper we propose an algorithm to modify the ordinary PCA biplot such that a timeseries PCA biplot is obtained where observations appear in a two-dimensional plot, but observations are ordered in time order from left to right. At the same time, the biplot is fitted with multiple biplot axes providing for reading off the values of the multiple timeseries. The iterative algorithm for constructing the timeseries PCA biplot is illustrated with a small subset of financial index data. After discussing some computational issues regarding the implementation of the timeseries PCA algorithm, the methodology is illustrated on the full financial index timeseries as well as an ecological dataset.

84

Generalising the molecular speed distribution of Maxwell

Prof Iain MacDonald[1], Dr Etienne Pienaar[1]
[1]University of Cape Town, Cape Town, South Africa

The molecular speed distribution proposed by Maxwell in 1860 is an early example of a probabilistic model in physical chemistry. The corresponding distribution of the squared speed is essentially a chi-squared distribution with three degrees of freedom, but chi-squared distributions (at least under that name) came later. To this day, most of the textbooks of physical chemistry (in their chapters on the kinetic theory of an ideal gas) do not mention chi-squared distributions and continue to make an assumption that Maxwell made: independence of the velocity components in the x, y and z directions. But that assumption has at times been challenged, and independence would fail at extremely low temperature or extremely high density.

We examine here the effect of relaxing that assumption. But we retain the assumption that the three velocity components are zero-mean normals with common variance. That is, we seek the distribution of the sum of squares (U) of three dependent zero-mean normals with common variance--- or the distribution of √U. A closed-form expression for the probability densities seems difficult to find, but there is a neat representation of U as a certain linear combination of independent random variables. That representation can be used to find (numerically) the p.d.f. of U or √U for a given value of ρ, the correlation between any two of the velocity components.

One consequence of the representation is that the mean of U is unaffected by the value of ρ, and hence the mean kinetic energy of a molecule is also unaffected. But the variance of U is an increasing function of ρ^2.

140

Identifying contributing factors to profile non-completing students in the faculty of natural sciences

Mr Edwin Mahlangu[1], Ms Xabsa Mohumed[1], Ms Kesia Phigeland[1], Dr Humphrey Brydon[1], Ms Khadija Parker[1], Prof Renette Blignaut[1]
[1]University of the Western Cape, Bellville, South Africa

This project identifies factors or variables that contribute to high non-completion rates among students enrolled in six programmes from the Faculty of Natural Sciences (FNS) at the University of the Western Cape (UWC). The data focused on students who registered for the first time at UWC from 2015 to 2019. Decision tree models were constructed for all programmes, with a 10-fold cross-validation to evaluate the models. The models showed that not having a bursary was the only key factor, masking all other factors from being identified by the decision tree. The results showed that the school quintile and the UWC admission point score were the next strongest factors after the bursary flag in almost all programmes. Profiles have been developed for students exhibiting characteristics associated with non-completion, providing valuable insights for the FNS to implement targeted interventions aimed at improving student success and retention.

12

Extreme value dependence analysis to bitcoin/us dollar and South African rand/us dollar exchange rates

Dr Katleho Makatjane[1], Mr Lethlogononolo Mosanawe[1], Dr Claris Shoko[1]
[1]University of Botswana, Gaborone, Botswana

This study proposes a dependence-switching quasi-vector autoregressive (VAR) copula model with a common trend. The aim is to examine extreme dependence and tail dependence for two exchange rate market statuses: appreciating currency and depreciating currency. The proposed model is estimated using the adjusted closing values of daily Bitcoin/US dollar and South African rand/US dollar exchange rates from January 02, 2015, to July 31, 2024. The extension of this non-stationary modelling in the literature is quite complicated since it requires specifications not only on how the parameter estimates change over time but also on those with bulk distribution components. The estimated autoregressive score dynamic allowed the copula parameters to promptly react to important systemic and time-varying indicators. For the two exchange rates, it is discovered that the dependence and tail dependence among the two markets as mentioned earlier statuses are symmetric in the negative correlation regime but asymmetric in the positive correlation regime. These findings add to previous literature and imply that a time-invariant copula framework may not be the best one to use when examining cross-market linkages.

80

Distribution-free generalised EWMA control charts using two-sample tests with application in froth flotation process

Miss Palesa Makena[1], Dr Majika Jean Claude Malela[1]
[1]University of Pretoria, Pretoria, South Africa

Classical control charts are mostly based on the assumption of normality and/or a specific probability distribution. When this assumption fails to hold, nonparametric charts are recommended. These charts are based on nonparametric tests such as the sign, signed-rank, precedence, exceedance, rank and Mann-Whitney tests. The latter is equivalent to the Wilcoxon rank-sum (WRS) W test, and it is considered to be the most powerful nonparametric test because of its high power of the test that enables it to likely find a significant difference between the means of two groups under the violation of the normality assumption. This paper develops a new distribution-free generalized exponentially weighted moving average (GEWMA) chart based on the WRS W statistic (denoted as W-GWMA chart). The new chart is designed using r smoothing parameters ($r \geq 1$) that carry a decreasing weight. The new chart is a special case of the EWMA chart where r= 1. The robustness and out-of-control (OOC) performance of the new chart are studied using the characteristics of the run-length distribution. It is observed that the W-GEWMA chart has very interesting in-control (IC) and OOC properties of the run-length distribution. The application and implementation of the new chart are demonstrated using real-world data from a froth flotation process.

165

Quantifying how fast South Africa's new car sales recovered from the COVID-19 pandemic using time series intervention analysis

Dr Tendai Makoni[1], Prof Delson Chikobvu[1]
[1]University of the Free State, Bloemfontein, South Africa

The South African automobile industry is critical to the economy and was significantly impacted by the COVID-19 pandemic. The objective of the paper is to quantify the influence of the COVID-19 pandemic on new car sales in the South African automobile industry. Time series intervention analysis offers a way to quantify and assess this impact. This type of analysis provides valuable information and insights for policymakers and stakeholders. The models allow for the effects of interventions on time series' normal behaviour, making it possible to quantify the impact of the COVID-19 pandemic. The SARIMA(0, 1, 1)(0, 0,2)12 model with an intervention component was confirmed as the best fit, based on the Akaike Information Criterion (AIC), root mean square error (RMSE), and the mean absolute error (MAE). In April 2020, there was an abrupt reduction (83.40% drop) in new car sales due to the COVID-19 pandemic and the subsequent economic lockdowns introduced in South Africa. The intervention impact was sudden but short-lived and did adversely affect the automobile industry. This underscores the importance of businesses having contingency plans, such as business interruption insurance during similar future disruptions. The recovery over time demonstrates the resilience of the automobile industry, indicating that despite the severe initial shock, the sector rebounded as economic conditions improved. This recovery trend signalled improved economic conditions as the economic lockdown conditions eased, and consumer confidence improved following the initial disruptions caused by the pandemic. The SARIMA-intervention model is a good tool for identifying intervention points and characterising the effects of these interventions or events on industries beyond car manufacturing. This can guide decisions on production schedules, inventory management, and strategic planning, ensuring operations are closely aligned with market conditions, especially as and when the industry begins recovery.

49

A rank-based EWMA TBEA control chart

Dr Majika Jean Claude Malela[1], Prof Fernanda Otilia Figueiredo[2], Prof Philippe Castagliola[3]
[1]University of Pretoria, Pretoria, South Africa, [2]University of Porto, Porto, Portugal, [3]University of Nantes, Nantes, France

Recently, considerable attention has been paid to the development of Time Between Events and Amplitude (TBEA) control charts. Almost all existing TBEA charts are of a parametric type. Parametric TBEA charts have the disadvantage of being very sensitive to deviations from the distributional assumptions and to the estimation of the process nominal parameters. This emphasizes the importance of developing nonparametric (or distribution-free) TBEA control charts. In this paper, a new distribution-free EWMA TBEA control chart based on the rank statistic, denoted as rank-based EWMA TBEA chart, for simultaneously monitoring the time interval between successive occurrences of an event and its magnitude is proposed. This chart is an extension of the Sign EWMA TBEA chart and uses a statistic close to the Wilcoxon Mann-Whitney statistic. The run length properties of the new TBEA chart are obtained by Markov chain techniques, and some numerical comparisons with other competing charts reveal its promising performance. An illustrative example is also provided to demonstrate the application and the implementation of the proposed TBEA control chart using real-world data.

107

Multivariate Bayesian small area estimation of health statistics indicators

Prof Samuel Manda[1]
[1]University of Pretoria, Pretoria, South Africa

The univariate Fay-Herriot model is commonly used to estimate reliable area parameter estimates of a public health variable produced from many surveys. In the era of multiple data being collected in an area, Multivariate Fay-Herriot (MFH) models that consider the correlation of several related health variables have become popular methods. We explore MFH Bayesian models and study the UNAIDS "95-95-95" targets, estimating adult recent HIV testing coverage, HIV prevalence, ARV uptake coverage, and HIV viral load suppression in small areas in Southern Africa. The four are associated with direct estimates from most surveys and are strongly correlated.

134

Trend analysis and determinants of violence against women in South Africa using VOCS 2013-2017 data

Mr Sonnyboy Manthata[1], Dr Lebogang Sesale[2], Prof Solly Seeletse[3]
[1]Sefako Makgatho University of Health Sciences, Pretoria, South Africa, [2]University of South Africa, Pretoria, South Africa, [3]Sefako Makgatho University of Health Sciences, Pretoria, South Africa

This study aims to improve the understanding of the prevalence and risk factors associated with Violence Against Women (VAW) in South Africa using the Victims of Crime Survey (VOCS) data collected between 2013 and 2017. By combining multi-year VOCS datasets, the study mainly focuses on females aged 16 years or older who had experienced assault or sexual offence incidents. Those who responded yes, were designated as being the victims of VAW. The victims were further classified into those incidents which were perpetrated by the partners and were identified as being victims of Intimate Partner Violence (IPV). The objective of the study included an analysis of the trends of VAW by province, investigating the relationship between VAW and factors such as age, marital status and education and an assessment of the impact of VAW and IPV on victims. The study's results revealed that age, marital status and province were significant predictors of VAW. Regarding IPV, age and province were significant predictors while highest level of education did have a significant effect. The odds ratios showed that women aged 44 years and older were less likely to be victims of VAW compared to females aged 25 to 34 years, while education had correlation with a reduced likelihood of VAW. Furthermore, the results of the study indicate that women in KwaZulu-Natal were at lower risk for IPV compared to those in Gauteng province, while married women were less vulnerable than their single counterparts. Future studies can be conducted to investigate the economic status of the victims and the reasons behind under-reporting of VAW and IPV in South Africa using victimisation surveys.

139

An analysis of new entrants in technical and vocational education and training colleges: 2022

Mr Sonnyboy Manthata[1], Ms Nthabiseng Tema[1], Ms Mmakgotso Ntsoane[3]
[1]Department of Higher Education and Training, Pretoria, South Africa

Since 2016, Technical and Vocational Education and Training (TVET) colleges are required submit detailed enrolment data through the Technical and Vocational Education and Training Management Information System (TVETMIS). However, inconsistencies in data accuracy, particularly in reporting new entrants, have posed challenges. To address these challenges, the Department of Higher Education and Training (DHET) developed a methodology in 2019 to calculate new entrants by merging TVET with NSC data from the Department of Basic Education (DBE) resulting in a publication of first report in 2020. The most recent data from 2022 revealed a significant decline in the number of new entrants, dropping from 243 534 in 2017 to 176 548 in 2022. This figure is also lower than the 194 714 new entrants recorded at public Higher Education Institutions (HEIs). The analysis shows a decrease in the proportion of new entrants in the total student population, from 50% in 2017 to 43.3% in 2022. Most new entrants (96.4%) in 2022 were aged 34 years or younger, while 3.6% were 35 years and older, a group that essentially requires Work Integrated Learning (WIL) to complete their National N Diploma. Sector Education and Training Authorities (SETA's) in the past only catered to students aged 15-35, leaving older students needing access to this crucial component which led to a policy shift allowing those older since 2022. Further analysis of the 2022 data reveals that over 30% of new entrants had completed Grade 12 in 2021, with 32.3% achieving Diploma pass and 21.3% a Bachelor's pass. The report also highlights gender disparities, particularly in National Certificate (Vocational) [NC(V)] programs, where females are predominantly enrolled in Office Administration, while males tend to pursue Engineering Studies. These findings suggest the need for targeted policy interventions to address gender imbalances and ensure equitable participation in vocational education programs.

18

Quantifying loss to the SA wholesale and retail industries using interrupted time series models

Mr Thabiso Masena[1], Mr Sandile Shongwe[1], Dr Ali Yeganeh[1]
[1]University of the Free State, Bloemfontein, South Africa

This study aims to estimate and quantify the actual amounts (in South African Rands) that the negative impact of the intervention effects of the COVID-19 pandemic had on the South African total monthly wholesale and retail sales using the seasonal autoregressive integrated moving average with exogenous components (SARIMAX) model. The SARIMAX model is supplemented with three approaches for interrupted time series fitting (also known as a pulse function covariate vector) which are: (i) Trial-and-error, (ii) quotient of fitted values and actual values, and (iii) a constant value of 1 throughout the intervention period. Model selection and adequacy metrics indicate that fitting a pulse function with trial-and-error approach produce estimates with the minimum errors on both datasets, so that more accurate loss in revenue in the economy can be approximated. Consequently, using the latter method, the pandemic had an immediate, severe negative impact on wholesale trade sales, lasting for 15 months (from March 2020 – May 2021) and resulted in a loss of R302 399 million in the economy. Moreover, the retail sales were also negatively affected, but for 8 months (from March 2020 to October 2020), with a 1-month lag or delay, suggesting that the series felt the negative effects of the pandemic one month into the intervention period and resulted in a loss of R85 298 million in the economy.

42

Application of extreme value theory to finance data

Mr Daniel Levy Mashilo[1]
[1]University of South Africa, Johannesburg, South Africa

This research proposal investigates Extreme Value Theory (EVT) applied to financial data, focusing specifically on the Johannesburg Stock Exchange (JSE). This research aims to address the limitations of traditional distributions, such as the normal distribution, which fail to adequately account for the negative skewness and excess kurtosis present in asset return distributions. Through the application of EVT, the investigation aims to ascertain the optimal parent distribution for modelling extreme return values, an attempt that is essential for precise risk assessment. The analytical framework will incorporate diverse EVT methodologies, including block maxima and peaks over threshold techniques, to assess their efficacy in capturing extreme events within financial datasets. Furthermore, the proposal will explore the synthesis of the Fréchet and Gumbel distributions to facilitate the development of a novel perspective on extreme values, specifically, those data points that are significantly elevated or diminished. Ultimately, this research aspires to yield significant insights regarding the dynamics of extreme financial phenomena, thereby enhancing risk management strategies within South African financial markets. The findings of this research will hold significant implications for investors and financial institutions attempting to navigate the complexities inherent in extreme market conditions.

161

The impact of clustering in randomised clinical trials: Scoping review and comparative statistical analysis

Ms Mikateko Mazinu[1], Prof S Manda[2], Dr T Reddy[3]
[1]South African Medical Research Council, Tygerberg, South Africa, [2]University of Pretoria, Pretoria, South Africa, [3]South African Medical Research Council, Durban, South Africa

Multicentre randomised controlled trials (RCTs) are widely used in phase 2 and 3 clinical research studies, particularly in the African region. These trials result in correlated observations within sites, which requires careful statistical analysis to ensure accurate results. Our scoping review, which was focused on HIV trials, found that many studies failed to account for this clustering in the design and analysis of endpoints, raising concerns about the validity of their findings. We found that under 20% of studies accounted for site-specific clustering. Within the studies which did account for clustering in the analysis, the most commonly applied approach was generalised estimating equations (GEE), followed by mixed effect models including site as a random effect. To enhance our understanding of clustering in clinical trials, we performed a series of analyses on a three-arm Phase 2a RCT of COVID-19 booster vaccines. We compared four statistical models: one that ignores clustering, and three that address clustering through different methods; site as a random effect, GEE, and regression with standard errors which allow for intragroup correlation. Our review and analysis highlight the importance of accounting for site clustering in multicentre RCTs, particularly in HIV research in Africa. While models that account for clustering provide reliable solutions, standardisation of reporting and application of these methods across trials is critical to avoid biased estimates and improve the reliability of trial results.

158

An analytical and empirical comparison of meta-analysis methods for individual participant binary data

Ms Abigail Mberi[1], Prof Samuel Manda[1]
[1]University of Pretoria, Pretoria, South Africa

Two-step and one-step methods are widely used in individual participant binary data meta-analyses. This paper presents both an analytical and a large-scale empirical comparison between methods. For the one-step methods, generalised linear mixed models (GLMMs) with log, logit, probit, and complementary log-log link functions were considered. Under the two-step meta-analysis, the standard fixed and random effects models were fitted. The methods were empirically compared in synthesising the effect of education and type of residence on cancer screening uptake in sub-Saharan Africa using data on 127 317 women aged between 15 and 49 years. The data was comprised of sixteen Demographic and Health Surveys (DHS) across ten Sub-Saharan countries. These surveys were obtained between 2009 and 2022.

39

Determination of predictors related to high blood pressure in South Africa using machine learning techniques

Dr Ruffin Mpiana Mutambayi[1], Mr Nhlonipho Mbhele[1], Mr Masimthembe Lala[1]
[1]University of Fort Hare, Alice, South Africa

The goal of the study is to determine which model is best suited for data analysis and display in the South African healthcare industry, with a focus on hypertension. To identify the predictors associated with high blood pressure in South Africa, various machine-learning techniques are applied throughout this study. The study used the South Africa Demographic and Health Survey 2016 dataset.
Machine-learning techniques such as random forest, K-NN, decision trees, Naive Bayes, and support vector machines were used in the classification. The dichotomous logistic model was used to identify the predictors associated with high blood pressure. With a precision of 0.87, the results show that random forest performed better in the classification process. Additionally, the dichotomous logistic model demonstrates that variables such as 'Age' (p-value < 2e-16; OR: 1.061537; CI: 1.0545341-1.06854023); 'Perception of own health' (p-value = 2.41e-07; OR: 0.677629; CI: 0.5299367-0.82532043); 'Wealth Index Combined' (p-value = 3.44e-11; OR: 1.351497; CI: 1.2624034-1.44059094); 'Region' (p-value = 2.43e-07; OR: 0.88203; CI: 0.8343747-0.92968563); 'Approach toward salt consumption' (p-value = 0.00282; OR: 1.082581; CI: 1.0305175-1.13464447); 'Last month received medical or dental care' (p-value = 8.79e-14; OR: 3.122288; CI: 2.8230624-3.42151315) and 'Respondent's perception of weight' (p-value = 0.01943; OR: 1.162578; CI: 1.0362502-1.28890593) are statistically significant and associated with the diagnosis of hypertension in South Africa, among other variables. The model's significance is confirmed by the ROC's accuracy of 0.84. With an estimation power of 87%, the Random Forest performed exceptionally well in the classification process. The accuracy level of the reduced dichotomous logistic model was 88%. The reduced model yielded additional findings, such as the identification of several covariates as predictors associated with hypertension in South Africa.

52

Proportion and risk factors associated with 'never tested for HIV' amongst women in Tanzania

Dr Sizwe Mbona[1], Prof Retius Chifurira[1], Dr Bonginkosi Duncan Ndlovu[1]
[1]Durban University of Technology, Durban, South Africa

Background: Despite several intensive interventions, the prevalence of Human Immunodeficiency Virus (HIV) remains a global health challenge affecting many individuals worldwide. Objectives: To assess the prevalence of 'never tested for HIV' and the risk factors associated therewith among women aged 15-49 years. Methods: The 2022 Tanzania Demographic and Health Survey (TDHS) data were used for this study. The variable of interest was reported never tested for HIV amongst women of reproductive age (WRA). A total of 15 254 WRA participated in the study. To identify the socio-factors associated with never tested for HIV, a survey logistic regression model was used due to the complexity of the sampling design. Analysis was performed at a 5% level of significance using STATA 16.0 software. Results: Of the 15 254 WRA that participated, 3 082 (20.2%) reported never being tested for HIV. The mean age of the participants was 29 (SD = 9.85) years. The odds of never being tested for HIV was 2.23 [odds ratio (OR) = 2.23; 95% Confidence Interval (CI) = 1.52 - 3.27] higher amongst women residing in rural areas as compared to their counterparts. Furthermore, factors such as level of education, breastfeeding, marital status and working status were associated with never being tested for HIV amongst WRA in Tanzania. Conclusion: This study identified the main factors influencing not testing for HIV amongst WRA, namely the level of education, breastfeeding, marital status and working status. To facilitate HIV testing amongst WRA, governments must develop intervention programmes that address the risk factors identified by this study.

20

Application of joint modelling and longitudinal latent modelling to antiretroviral adherence monitoring

Mr Campbell Mcduling[1], Ms Lauren Jennings[2], Prof Catherine Orrell[2], Prof Francesca Little[1]
[1]Department of Statistics, University of Cape Town, Cape Town, South Africa, [2]Center for Adherence and Therapeutics, Desmond Tutu Health Foundation, Cape Town, South Africa

Monitoring adherence to antiretroviral therapy is critical in managing the HIV/AIDS epidemic, enabling rapid identification of non-adherent individuals and subsequent intervention. Joint modelling (JM) methodologies can examine the association between virologic outcomes and mechanisms for monitoring longitudinal adherence. Once valid monitoring tools are identified, behavioural insights can be extracted from longitudinal data through analyses based on finite mixture modelling. Group-based trajectory models (GBTMs) offer a promising approach to identify latent classes of adherence behaviour from longitudinal adherence monitoring data.

We applied semi-parametric survival models with recurrent events to analyse the association between time to viral non-suppression and time-varying adherence data, extending this via the JM framework, to 12 months of electronic monitoring (EM) device data from a 24-month prospective observational study. The cohort consisted of 250 virally-suppressed people living with HIV in the Western Cape region of South Africa. Following this, a GBTM was used to identify a finite number of heterogenous adherence trajectories. We then applied a Kaplan-Meier survival analysis on the sample, after stratifying by adherence profile.

The JM results suggested that each additional missed dose in the preceding 30 days was associated with an estimated conditional hazard ratio of 1.81. The GBTM analysis revealed five distinct adherence trajectories: 1) stable and excellent adherence (19% of participants), 2) stable and acceptable adherence (26%), 3) stable and poor adherence (17%), 4) slowly deteriorating adherence (18%), and 5) rapidly deteriorating adherence (20%). The Kaplan-Meier analysis showed these subgroups exhibited markedly different viral suppression outcomes, with the two deteriorating adherence groups showing a much faster average decline in the probability of viral suppression over time.

These analysis approaches provide valuable tools for assessing the validity of adherence monitoring data and for identifying heterogeneous patterns of longitudinal ART adherence that have important implications for clinical management and viral outcomes.

89

Logistic regression analysis to identify the determinants of concurrent sexual partnership among Kenyan women

Dr Tshaudi Motsima[1], Mr Banele Mdakane[1], Ms Thelma Maunye[1]
[1]Tshwane University of Technology, Pretoria, South Africa

The practice of concurrent sexual partnerships is not unusual despite the almost universal encouragement of the norm of fidelity in marriage. Married women enter concurrent sexual partnerships for various and the practice of concurrent sexual partnerships reduces marital quality and is one of the reasons for collapsing the marriage. It exposes married couples to the risk of contracting sexually transmitted infections, human immunodeficiency virus, and acquired immunodeficiency syndrome. The purpose of this study is to examine the factors associated with concurrent sexual partnerships among married women in Kenya. The 2022 Kenyan Demographic and Health Survey (KDHS) data were used. Logistic regression model was employed to identify the determinants of concurrent sexual partnership among married Kenyan women. It was found that education level, staying with husband, religion, place of residence, condom use in recent sexual activity, and genital sore were determinants of concurrent sexual partnership among married Kenyan women. Less educated women, women who do not stay with their husbands, urban-based women, women who used condoms in their recent sexual activities, and women who had genital sores were more likely to be involved in concurrent sexual partnerships than their counterparts. Muslim women were less likely to enter concurrent sexual partnerships than their counterparts.

79

Survival analysis of time-to-credit default in the presence of time-varying covariates

Mr Lusanda Mdhlalose[1]
[1]University of the Witwatersrand, Johannesburg, South Africa

This research investigates the prediction of time-to-credit default in mortgage loans using both traditional statistical methods and machine learning models in survival analysis. We evaluated the Extended Cox model, including its penalized variants (Ridge and Lasso regression), alongside the Left Truncated and Right Censored Conditional Inference Forest (LTRCCIF). The models demonstrated strong dependence on time-varying covariates like loan-to-value ratio, outstanding loan balance, interest rate, and some fixed covariates such as FICO (credit) score, and real estate type. The models were fitted and evaluated across 12, 24, 36, 48 and 60 month periods. The Extended Cox model in discrimination with C index had values ranging around 0.61 and 0.64 across the different time periods. The penalized Cox models improved performance on some time periods, partially due to inherent feature selection using regularization. Ridge regression outperformed the unpenalised model in discrimination with C-Index values ranging between 0.56 and 0.70, while Lasso regression offered a more compact model with C-Index values ranging between 0.56 and 0.65. The LTRCCIF model showed exceptional risk discrimination, maintaining high C-Index values across all time points, ranging from 0.95 to 0.98. In terms of calibration, the Integrated Brier Scores (IBS) for the Extended Cox ranged around 0.006 and 0.12 while its penalized variants ranged between 0.007 and 0.13. The LTRCCIF model ranged between 0.005 and 0.10. These values suggest variable calibration performance between the models depending on the time period observed. Overall, machine learning models, particularly LTRCCIF, showed superior performance in risk differentiation, while traditional methods excelled in incorporating a wide variety of covariates in feature importance analysis.

15

Use of some important statistical methods in electrical energy generation and their applications

Dr Vincent Micali[1]
[1]Stats4buz (Pty) Ltd, Hout Bay, Cape Town, South Africa

The generation of Electrical Energy is an essential component of a Country GDP. Forecasting the energy requirements (in MWh) and predicting the maximum demand (in MW) are critical elements for the sustainability of the Country's economy. This presentation provides a statistical methodology in forecasting, predicting, performance modelling according to strategies, decided upon by a generating Utility, to achieve focused targets. Monitoring (data acquisition and validation), Evaluations and Performance Management constitute a vital information source for the decision makers. Here, the dynamics of the information provided within the pre-defined risks is shared. As any business should adhere to the AARA tenets (Accessibility, Availability, Reliability, Affordability, in order of importance), in their product(s), Electrical Energy production and delivery is no different. These Tenets need dependable statistical foundations for the decision makers. Methods, processes and Utility internal dynamics are depicted here. Performance Management processes, measuring instruments and contractual obligations are also described. An example of a vertically integrated Utility forms the cornerstone of this topic. It's achievements as a practical application are also imparted, with preservation of confidentialities. The intent of this presentation is to provide, as high-level view in a Utility, the position of a Statistician as a competent person in this Industry.

85

Divergence-based approach in bivariate tail dependence coefficient estimation

Dr Richard Minkah[1,3], Prof Abhik Ghosh[2], Prof Tertius de Wet[3]
[1]University of Ghana, Accra, Ghana, [2]Indian Statistical Institute, Kolkata, India, [3]Stellenbosch University, Stellenbosch, South Africa

This study addresses the estimation of the bivariate tail dependence coefficient from a robustness perspective, merging the principles of robust statistics with those of extreme value statistics. Traditional estimators, such as those based on maximum likelihood or moments, are highly affected by outliers in the data. To tackle this issue, we introduce a robust estimator that minimizes the density power divergence under appropriate model assumptions. The robustness of this estimator is analysed using classical influence function methods. Additionally, we demonstrate the estimator's effectiveness through a comprehensive empirical study involving various significant bivariate extreme value distributions. Ultimately, the proposed estimator is applied to estimate the tail dependence coefficient for a real-world dataset related to workers' compensation.

78

Analysing exercise-associated muscle cramping in ultramarathon runners using predictive modelling

Ms Xabsa Ahmed Mohumed[1], Dr Retha Luus[1], Ms Tayla Wannenberg[1], Mrs Esme Jordaan[2]
[1]University of the Western Cape, Bellville, South Africa, [2]South African Medical Research Council, Parow, South Africa

The aim of this study is to predict the prevalence of exercise-associated muscle cramping (EAMC) using data from the 2022 Comrades Marathon. Previous research by MacMillan et al. (2024) identified a history of EAMC (hEAMC) prevalence of 14.4%, suggesting a possibly significant target class imbalance. The authors used log-binomial regression to identify risk factors associated with hEAMC but did not account for the target class imbalance which could pose a challenge for this modelling technique.

In this project, various sampling methods and modelling techniques are explored to address the class imbalance before predicting hEAMC. The log-binomial regression serves as the baseline model. For comparison to the baseline model, a decision tree, random forest, gradient boosting, and neural network are also considered. This modelling strategy will be applied to both the imbalanced and balanced datasets, obtained through oversampling, e.g., synthetic minority oversampling technique (SMOTE), and under-sampling. Using a hold-out dataset, each model's predictive accuracy will be assessed with metrics such as AUC-ROC, precision, recall, and F1 score. Model sensitivity to random partitioning will be accounted for by using five random partitions before selecting the champion model. The goal is to find the best model which provides an insight into the most important factors predicting hEAMC. It is envisaged that this research will enhance the understanding of hEAMC, offer practical recommendations to optimise athlete performance and reduce muscle cramping incidence in long-distance running events.

References

MacMillan, C., Sewry, N., Schwellnus, M., Boulter, J., Dyer, M., & Jordaan, E. (2024). Sex, training variables, history of chronic disease, and chronic injury are risk factors associated with a history of exercise-associated muscle cramping in 10,973 ultramarathon race entrants: A safer XXXVIII study. The Journal of Sports Medicine and Physical Fitness. https://doi.org/10.23736/S0022-4707.24.15842-

94

Comparing the power of multivariate test statistics for three-factor interaction in a 3-way contingency table

Ms PB Mokoena[1]
[1]University of South Africa, Florida, South Africa

Background: Analysing interactions in contingency tables is essential for fields like epidemiology and public health. Traditional multivariate tests, including Pearson's Chi-square, the Likelihood Ratio Test (LRT), and the Product Moment Chi-square (PMC), have varying effectiveness based on data sparsity. Log-linear analysis, using Maximum Likelihood Estimation (MLE), is recommended for larger datasets. This study compares the power of these methods in detecting three-factor interactions.

Objective: To compare the power of Pearson's Chi-square, LRT, PMC, and log-linear analysis in identifying three-factor interactions in 3-way contingency tables, particularly under varying data sparsity conditions.

Methods: Death records from 2009 to 2018, categorised by year, gender, and age group, were analysed. Simulations were used to assess the power of each test, and log-linear analysis was tested for handling sparse data.

Results: Test performance varied with data sparsity. Log-linear analysis showed strong results for larger datasets, while traditional methods were less effective with sparse data.

Conclusion: Log-linear analysis is more suitable for large, sparse datasets due to MLE. Traditional tests like Pearson's Chi-square and LRT showed reduced power when faced with sparsity, highlighting the need to consider sample size and data characteristics in multivariate analyses.

9

Estimation of covariance function of a stationary ARMA process

Dr Wessel Moolman[1]
[1]Akademia, Centurion / Die Hoewes, South Africa

The two estimators of the lag k covariance function of a stationary process are the sample covariance function with either n (series length) or n-k as divisor. The talk is about comparing the MSE's (mean squared errors) of the two estimators for a variety of ARMA models [AR(1), AR(2), MA(1), MA(2) and ARMA(1,1)] and to opt for the estimator with the smaller MSE. Comparisons of MSE's for the two estimators for medium and large samples will be made.

81

Predictors of emotional and physical abuse towards Kenyan men: a logistic regression analysis

Dr Tshaudi Motsima[1]
[1]Tshwane University of Technology, Pretoria, South Africa

Gender-Based Violence is a universal public health concern and a violation of human rights. Its effects include emotional-trauma and mortality. Globally, there is little research focusing on GBV against men because many studies are subjective towards women as victims. Logistic regression was applied to the 2022 Kenya DHS data to establish the driving factors of emotional and physical abuse against men. Findings showed that education, alcohol, marriage type, number of women the man fathered children with, age at first cohabitation, wealth and place of residence were significant driving factors of abuse towards men. Age at first cohabitation, wealth and place of residence were significant driving factors of physical abuse by women towards their husbands/partners. The government should design support groups to help men and train police officers to screen for violence when men report it. Legal and policy framework ought to be gender/sex neutral. The United Nations should include violence against men in its programmes, such as the SDGs. Men should improve their economic standings by findings jobs or try to make other means to generate income. They should reduce alcohol intake and delay entry to marriage/cohabitation.

95

An illustration of gender differential item functioning analysis in mathematics from national benchmark tests

Mrs Precious Mudavanhu[1]
[1]University of Cape Town, Cape Town, South Africa

The National Benchmark Tests (NBT) Project assesses students' competencies in Academic Literacy (AL), Quantitative Literacy (QL), and Mathematics (MAT) to identify writers in need of academic support, determine appropriate programme placements, and inform curriculum development. As high-stakes assessments, NBT plays a critical role in shaping educational pathways, making it essential to ensure that the tests are fair, reliable, and free from bias. To keep tests free from bias, various statistical and psychometric tools, such as differential item functioning (DIF) analysis, are used to identify and eliminate problematic test items that may unfairly advantage or disadvantage certain groups of students. DIF analysis helps detect items where performance differences between subgroups are not solely attributable to the abilities being measured. This study illustrates the application of the DIF method using the two-parameter Item Response Theory (2PL IRT) model in Xcalibre in evaluating NBT MAT test items for potential bias related to gender. The 2023 administration of the MAT test provided score data of at least 14 test forms that had more than 950 writers. To obtain accurate estimates for the parameters of the item in IRT, the theory requires a large sample size (around 1000). All the MAT test forms consist of 60 items which includes 8 anchor items. The data were subjected to DIF analysis through a 2PL IRT procedure. The results indicated less than 2 items on average were flagged for DIF and the items were biased against the female group. The findings from suggested that the NBT MAT test does not extensively suffer from DIF. It is recommended that the test developers review the flagged items and possibly remove them from the item bank.

38

Analysis of predictors related to diagnosis of hypertension correlated with heart attacks in South Africa

Dr Ruffin Mpiana Mutambayi[1], Mrs Natalie Benschop[2], Prof Retius Chifurira[2]
[1]University of Fort Hare, Alice, South Africa, [2]University of KwaZulu-Natal, Durban, South Africa

High blood pressure is increasing worldwide, and South Africa is no exception. High blood pressure can be fatal if it causes serious events, such as a heart attack. Therefore, it is critical to understand not only the etiology of hypertension but also why some people with hypertension have heart attacks while others do not. The study used the South Africa Demographic and Health Survey 2016 dataset. The random forest technique and polytomous logistic model were used to identify risk factors related to the association between heart attacks and high blood pressure to achieve the study's goals and objectives. The reduced polytomous model was generated, and it was statistically significant (p-value < 2.2e-16) and produced a residual deviation of 2797.787 and an AIC of 2863.787.
The findings from the reduced polytomous model show that predictors such as 'Use medication regularly prescribed by a doctor/nurse' (p-value <2.2e-16); 'Age in 5-year groups' (p-value <2.2e-16); 'Region' (p-value = 0.0001891); 'Wealth index for urban/rural' (p-value = 0.00369); 'Has a doctor or nurse told that you had TB' (p-value = 0.009903); 'Number of wives/partners' (p-value = 0.05336); 'Last 12 months have woken up with tightness in the chest' (p-value = 0.01189); 'Last 12 months respiratory problems' (p-value = 0.03488); 'Ever worked in a place exposed to smoke, dust, fumes' (p-value = 0.0003186) and 'Respondent own perception of weight' (p-value = 0.008155) are related to the diagnosis of hypertension combined with heart attacks. The reduced polytomous logistic model had an 89.2% prediction power and performed better than the full model. Additional findings from the reduced model included the identification of multiple covariates as predictors connected to hypertension-related heart attacks.

173

Distributions of wet and dry spells

Ms Nothabo Ndebele[1]
[1]University of The Witwatersrand, Johannesburg, South Africa

Daily rainfall observations were used to characterise consecutive wet and dry days (spells) for the winter rainfall region in the Western Cape. Probability models in the geometric family of distributions were explored in terms of their ability to describe wet and dry spells. The differences in the distribution parameters for the fitted models were compared for the annual spells and wet season spells that occur during the months starting from April to the end of October. Rainfall distribution and quantity in this region is highly variable due to topographic influence and direction of rain bearing systems, thus the distribution parameters were compared across weather stations. Data from 8 weather stations with more than 50 years of available data were used in the study. Mean spell lengths ranged between 1 and 3 days for wet spells and between 4 and 6 days for dry spells during the wet season.

59

A nonparametric estimation of cumulative incidence functions in the presence of cured subjects

Dr Bonginkosi Duncan Ndlovu[1], Dr Sizwe Vincent Mbona[1]
[1]Department of Statistics, Durban University of Technology, Durban, South Africa

Competing risk data arises in survival analysis experiments that have multiple modes of failure. In some instances, this data may come with a sizable proportion of cured subjects, i.e., data may be a mixture of cured and uncured subjects. In such situations, it is important to apply analysis methods that take into account the mixed nature of the data because the standard analysis methods tend to produce biased estimates. The methods for analysis of mixed competing risks data that have been suggested to date rely on some form of an Expectation Maximization (EM) algorithm (Dempster et al., 1977) for estimation of the relevant quantities. We present an alternate non-parametric method for analysis of mixed competing risks data, this method combines the method proposed by Maller and Zhou (1992) for estimating the cured proportion non-parametrically, and the vertical cure model (Nicolaie et al., 2019). The advantage of the proposed method lies in the estimation procedure which circumvents the application of an EM algorithm.

14

The GARCH-EVT – Gumbel copula approach to quantifying portfolio diversification effects

Mr Thabani Ndlovu[1], Prof Delson Chikobvu[1]
[1]University of the Free State, Bloemfontein, South Africa

The study uses the hybrid model of exponential generalised auto-regressive conditional heteroscedasticity (eGARCH)- extreme value theory (EVT) - Gumbel copula model to quantify the diversification effects (DE) of an equally weighted portfolio that contains Bitcoin and the South African Rand. Firstly the eGARCH(1,1) model is fitted to the returns data to capture volatility cluster. Secondly, the mixture model of GPD-Gaussian kernel-GPD is fitted to the standardised residuals from an eGARCH model. The generalised Pareto distribution (GPD) is fitted to the tails of the standardised residuals while the Gaussian kernel is used in the central parts of the data set. The GPD is used to characterise the heavy tails and hence extreme risk in the returns of the two assets. The Gumbel copula, an extreme value copula, was preferred because of its versatile ability to model different forms of dependencies. When $\alpha = 1$, the Gumbel copula reduces to an independent copula, and when $\alpha \rightarrow -\infty$, a rotated Gumbel copula can be used to model the margins. The Gumbel copula parameter is $\alpha = 1.009$, implying the currencies are independent. The positive Extreme Value Indices implying that the marginals follow a heavy-tailed distribution. At 90%, 95%, and 99% levels of confidence, the portfolio DE using VaR quantities are 30.26%, 28.45% and 28.04% respectively. While the portfolio DE using expected shortfall is 26.80%, 25.36% and 21.23% respectively. This implies that there is a significant reduction of potential losses (diversification benefits) in the portfolio compared to the risk of the simple sum of single assets. These results can be used by fund managers, risk practitioners, and investors to decide on diversification strategies that can help reduce their risk exposure. The findings suggest that there is a reduction in risk if one invests in a portfolio of both Bitcoin and the Rand.

125

Designing an optimal survey sample with predetermined sample sizes for subgroups

Dr Ariane Neethling[1], Mr Francois Neethling[2]
[1]Independent Statistical Consultant, Bloemfontein, South Africa, [2]Independent Statistical Consultant, Cape Town, South Africa

Many research studies rely on survey sample data, whether for academic or practical business purposes. Effective sample design is fundamental to successful research, as a poorly designed sample can lead to insufficient data that neither survey weights nor statistical methods can adequately correct. While simple random sampling is often idealised, most surveys require more complex sampling techniques to achieve a representative and optimal sample. This presentation introduces an innovative approach to designing a sample with predetermined sizes for various stratification variables, such as provinces, urban/rural areas, and population groups. A unique method for determining interlaced sample sizes across these strata will be presented, ensuring optimal allocation while meeting the predetermined subgroup sizes.

67

Optimal grid selection in spatial statistics

Ms Jamie-Lee Nel[1], Dr René Stander[1], Mr Kabelo Mahloromela[1], Prof Inger N Fabris-Rotelli[1]
[1]University of Pretoria, Pretoria, South Africa

Various types of spatial analysis and modelling require grids, for example spatial similarity tests, spatial homogeneity tests, crime forecasts, urban land use identification, landslide susceptibility models, groundwater modelling and digital elevation models. Thus, the grid choice in spatial statistics must be carefully considered as this choice affects all subsequent calculations and modelling. There are no guidelines on how to choose the size and shape of the grid, with the choice usually made based on the application field and guided by expert advice. Further, when conducting tests for homogeneity and similarity, for example, the cells need to be sampled to avoid the natural occurrence of spatial dependency of the data in each grid cell, since these tests rely on an assumption of independence. The objective of this research is to define the optimal grid size for different spatial data types as well as to use experimental design to optimally sample grid cells.

163

Parametric analysis of multistate survival modelling for birth parity transitions in rural South Africa

Ms Thambeleni Portia Nevhungoni[1], Dr Tarylee Reddy[1], Prof Samuel Manda[2], Prof Din Chen[2]
[1]South African Medical Research Council, Pretoria, South Africa, [2]University of Pretoria, Pretoria, South Africa

Birth parity intervals of women are strongly associated with maternal and neonatal morbidity in most low and lower-middle-income countries where fertility rates are high. Successive birth parity intervals have been modelled using time-to-event statistical regression techniques, sometimes extended to include women's parametric shared frailty effects. However, these models cannot fully incorporate the sequential parity events, where the parity number could be considered parity states, starting at 0 parity. Multi-state survival models allow direct and simultaneous modelling of multiple birth parity pathways where women are observed in various parity states. We employ parametric multistate survival models with transition-specific distributions to analyse birth parity transitions collected in Health and Socio-Demographic Surveillance Systems in Rural South Africa.

88

Understanding macroeconomic factors' influence on South African maize production and food security: VECM analysis

Ms Cynthia Boitumelo Ngwane[1,2], Mr Kajingulu Malandala[1]
[1]University of South Africa, Florida, South Africa, [2]Agricultural Research Council, Pretoria, South Africa

As the eighth largest maize producer in the world, South Africa holds a significant position in global and regional agricultural markets. Maize is the country's second-largest crop after sugarcane and is essential for local consumption, contributing to food security. However, South Africa's food security landscape is complex, affected by several socio-economic factors that shape both production and access to staple foods. This study explores the relationship between maize production and socio-economic variables such as maize price volatility, population growth, unemployment, the consumer price index, and household disposable income, and assessing how these factors influence food security in both the short and long term. The relationship between food production and food security is underpinned by the four pillars of food security: availability, access, utilisation, and stability. In this context, maize production is a proxy for food availability, and any fluctuations in production can have direct consequences for national food security. Using a Vector Error Correction Model (VECM), it becomes evident that socioeconomic factors affect maize production differently in the short and long term. In the long run, factors like white maize prices and household disposable income negatively influence maize production, while yellow maize prices, population growth, consumer price index, and unemployment have a positive effect. In the short run, however, the primary factors that negatively affect maize production are population growth and household disposable income.

98

Application of marginal theory for variable selection in partially linear models

Dr Mina Norouzirad[1], Dr Ricardo Moura[1], Prof Mohammad Arashi[2], Dr Filipe Marques[3]
[1]Center for Mathematics and Applications (NovaMath), NOVA School of Science and Technology (NOVA FCT), Caparica, Portugal, [2]Department of Statistics, Faculty of Mathematical Sciences, Ferdowsi University of Mashhad, Mashhad, Iran, [3]Center for Mathematics and Applications (NOVA Math) and Department of Mathematics, NOVA School of Science and Technology (NOVA FCT), Caparica, Portugal

One semiparametric regression model that allows for the flexible capture of complicated relationships within datasets is a partially linear model (PLM). PLMs integrate both linear and nonlinear predictors. Nevertheless, the presence of low-variance predictors provides a substantial challenge, as it complicates variable selection and degrades model performance. A new marginal theory-based method is suggested for addressing this issue, and instead of the L1 norm used in the LASSO estimator, a new penalty function is employed. This motivates the development of Marginalized LASSO, a novel penalized estimator designed for PLMs. This estimator enhances its practicality and simplicity of application due to its closed-form nature and also it is a competitor to the LASSO estimator in PLMs. Both estimators are evaluated in a comprehensive simulation study to see how well they perform in estimating and prediction, and then their predictive power is investigated with a real dataset that focuses on luxurious house price predictions in King County, USA. In both simulated and real-data situations, the results show that the Marginalized LASSO estimator consistently performs better than the LASSO estimator, highlighting its practical significance and potential impact.

19

Profiling students at risk of dropout at a university in South Africa

Dr Piet Ntema[1]
[1]University of Limpopo, Polokwane, South Africa

Student dropout is a significant concern for university administrators, students and other stakeholders. Dropout is recognised as highly complex due to its multi-causality, which is expressed in the existing relationship in its explanatory variables associated with students, their socio-economic and academic conditions, and the characteristics of educational institutions. This paper reports on a study that drew on university administrative data to build a profile of students at risk of dropout from 2008–2018. The study employed a data mining technique in which predictors were chosen based on their weight of evidence (WOE) and information value (IV). The selected predictors were then used to build a profile of students at risk of drop-out. The findings indicate that at-risk students fail more than four modules in a year with a participation average mark of 43% or less and have joined the university in the second academic year. It is suggested that universities put measures in place to control and prevent students who carry over four or more modules from adding modules to their registration until the failed modules are passed.

154

Statistical data analysis of a multidimensional binary data using chi-square tests and correspondence analysis

Ms Nombasa Ntushelo[1], Dr Itumeleng Matle[1], Mr Thabo Nkabinde[2]
[1]Agricultural Research Council, Stellenbosch, South Africa, [2]Agrispace, Cape Town, South Africa

A presentation about 2 well known statistical methods used in the data analysis of a categorical data. Chi-Square tests (Ott and Longnecker, 2016) and a correspondence analysis (Greenacre, 2007) have so much in common but different statistical methods. Chi-Square tests are used to make inferences about population proportions. For instance, a Chi-Square test of independence to test for associations between 2 categorical variables. Correspondence analysis (CA) is an exploratory multivariate statistical technique used to show associations between levels of 2 categorical variables in a 2-dimensional plot. The purpose of the presentation is to show the data analysis of a multidimensional binary data using the 2 statistical methods together. The multidimensional data was generated from a prevalence study at Onderstepoort Veterinary Institute of Agricultural Research Council. The data was made up of 274 chicken samples categorized by 7 salmonella serovar types. Bacterial isolates of 274 chicken samples were tested against 16 antibiotics, 21 resistance genes and 15 virulence genes. The observed value was binary. The research questions were to know the antibiotic resistance of bacterial isolates, which resistance gene associated with it and which virulence gene associated with the bacterial isolate. Chi-Square tests were performed using Frequency Procedure (PROC FREQ) of SAS statistical software. Correspondence analyses were performed using XLSTAT software. Discussion of Chi-Square test of homogeneity results including table of frequencies and table of column percentages. A combined discussion of Chi-Square test of Independence and correspondence analysis results including table of frequencies and table of row percentages. The correspondence analysis results include 3 types of CA plots namely a symmetric plot, asymmetric column plot and asymmetric row plot. In conclusion, the combination of the results from the 2 methods were used to answer the research questions.

90

The analysis of the cosmological parameters using maximum likelihood estimator and chi-square

Miss Sinenhlanhla Nxumalo[1], Prof Syamala Krishnannair[1]
[1]University of Zululand, KwaDlangezwa, South Africa

This study explores the estimation of key cosmological parameters, specifically the Hubble constant (H_o) and the matter density parameter (Ω_m), using two statistical methods: Maximum Likelihood Estimator (MLE) and Chi-square (χ^2) analysis. The methods are applied to two observational datasets of Hubble parameter (H(z)) measurements across different redshifts, with one dataset containing 57 data points and the other 31. The results show that the larger dataset with 57 measurements produces more constrained parameter estimates, evident from narrower contour plots and lower uncertainty. These findings align with the ΛCDM (Lambda Cold Dark Matter) model, reinforcing its validity while also contributing to the ongoing discourse surrounding the Hubble tension, the discrepancy in measurements of the universe's expansion rate. The work underscores the importance of both the quality and quantity of observational data in cosmological parameter estimation.

44

Cytokine profiles as predictors of HIV incidence using machine learning survival models and statistical interpretable techniques

Ms Sarah Ogutu[1], Dr Mohanad Mohammed[1], Prof Henry Mwambi[1]
[1]University of KwaZulu-Natal, Pietermaritzburg, South Africa

Cytokine profiles are increasingly recognised as potential predictors of HIV incidence due to their role in immune regulation and inflammatory responses. Traditional statistical methods for time-to-event data, such as the Cox proportional hazards (PH) model, may have limitations in handling high-dimensional data and non-linear relationships between predictors. Machine learning (ML) survival models may address these issues, including violating the PH model assumption. Survival support vector machine (SSVM) and random survival forest (RSF) models using the change or mean in cytokine values as predictors were used to investigate the association of HIV incidence and cytokine profiles, evaluate variable importance, and assess predictive accuracy using the concordance index (C-index) and integrated Brier score (IBS). We interpreted the models' predictions using Shapley additive explanations (SHAP) values. The RSF models exhibited superior performance over the SSVM models, and the difference covariate model outperformed the mean covariate model. The best C-index for the SSVM model was 0.7180 under the difference covariate model. For the RSF, it was 0.8801 under the difference covariate model using the log-rank split rule. Key cytokines identified as positive predictors of HIV incidence included TNF-A, BASIC-FGF, IL-5, MCP-3, and EOTAXIN, while a broader set of 29 cytokines were negative predictors. Baseline variables like frequency of condom use, treatment, number of partners, and sexual activity were also strong predictors. This study underscored the potential of cytokine profiles in predicting HIV incidence and demonstrated the competitiveness of RSF models for analysing high-dimensional, time-varying data over SSVM.

122

Stochastic modelling on rainfall variability in northern Nigeria

Prof John Olaomi[1], Mr James Ehimony[1,2]
[1]Department of Statistics, University of South Africa, Florida, Johannesburg, South Africa,
[2]Department of Statistics, Kogi State Polytechnic, Lokoja, Nigeria

The study explored the stochastic process in modelling the distribution of rainfall from some selected stations in Northern region of Nigeria. Continuous-time Markov chain was used to determine the level of persistence based on its relative frequencies. The rainfall data were collected from the stations spread across the Sahel, Sudan and Guinea savanna. An Exponential probability distribution was used to model the distribution of rainfall intensity after clustering the average rainfall experienced in all the stations. The extreme rainfall and the intensity of dryness, over the recorded period, across all the stations were observed. It was found that the change in climatic conditions of each station depends on the amount of rainfall experienced annually. This study helps in simulating the likely rainfall to be experienced in each station of the Northern Nigeria. It will also assist the Meteorological Agencies to make short term probabilities prediction for aviation and agricultural production sectors.

118

A contaminated negative binomial model for count health data

Mr Arno Otto[1]
[1]University of Pretoria, Pretoria, South Africa

In medical and health research, investigators are often interested in countable quantities such as hospital length of stay or the number of doctor visits. Poisson regression is commonly used to model such count data, but this approach can't accommodate overdispersion — when the variance exceeds the mean. The negative binomial (NB) distribution (NB-D), and by extension, NB regression, provide a well-documented alternative; however, real-data applications present additional challenges that must be considered. Two such challenges are: i) the presence of (mild) outliers that can influence the performance of the NB-D, and ii) the availability of covariates that can enhance inference about the mean of the count variable of interest. To jointly address these issues, we propose the contaminated NB (cNB) distribution that exhibits the necessary flexibility to accommodate mild outliers. This model is shown to be simple yet elegant in application, as well as intuitive in interpretation. In addition to the parameters of the NB-D, our proposed model has a parameter describing the proportion of mild outliers and one specifying the degree of contamination. Then, to allow available covariates to improve the estimation of the mean of the cNB distribution, we propose the cNB regression model. An expectation-maximization algorithm is outlined for parameter estimation, and its performance is evaluated through a parameter recovery study. Additionally, a sensitivity analysis is conducted to investigate the performance of our proposed models. The effectiveness of our model is demonstrated on two health datasets, where it outperforms well-known count models.

Investigating the robustness of clustered point pattern simulation

Miss Amy Pieters[1], Dr Rene Stander[1], Prof Inger Fabris-Rotelli[1], Mr Kabelo Mahloromela[1], Dr Renate Thiede[1]
[1]University of Pretoria, Pretoria, South Africa

Spatial point pattern analysis considers the arrangement of spatial locations and whether there is an underlying pattern. In this research, we specifically consider clustered point patterns, a spatial point pattern where the points attract each other. Investigating clustered point patterns is interesting as it can help highlight problem areas in the simulation of such point patterns. Various cluster point pattern types will be simulated and the robustness of the simulation examined by using functions available in the spatstat library in R. The point patterns are simulated, then a point process model is fitted to the data and a new point pattern is simulated from the fitted model. The simulated and resimulated point patterns are compared using the K-function and Kolmogorov-Smirnov tests. We conclude with a proposed methodology to use when simulating or fitting clustered point pattern data.

16

Clustering and classifying global food insecurity index and crop production using machine learning algorithms

Mr Jaden Pieterse[1]
[1]Sol Plaatje University, Kimberley, South Africa

Food insecurity continues to impact millions globally, prompting researchers to investigate its underlying causes. Despite efforts to address this issue, gaps remain in understanding how socio-economic and climatic factors influence major crop production, contributing to food insecurity. Various statistical and machine learning methods have been employed to tackle this problem. However, while machine learning approaches may yield accurate predictions, they often lack interpretability compared to traditional statistical models, highlighting the need for alternative models that are both accurate and interpretable. This study aims to compare the performance of the K-nearest neighbour (KNN), Random Forest (RF), K-Means algorithm, and Gaussian Mixture Model (GMM) in classifying and clustering global food insecurities and crop production. KNN and RF will be used to classify countries based on the food insecurity index and crop production (maize, wheat, and rice). The performance of these methods will be evaluated using metrics such as the Receiver Operating Characteristics (ROC) and Area Under Curve (AUC). Meanwhile, the K-Means algorithm and GMM will be employed to cluster countries, with their performance assessed using the Silhouette Coefficient and Dunn's Index. The study will utilize features such as the consumer price index, agricultural land, population, and climatic factors such as CO2 emissions, temperature, and precipitation, collected from multiple online sources for 2022, to classify and cluster the food insecurity index and crop production. The goal is to identify more robust and interpretable models that can pinpoint socio-economic and climatic factors affecting crop production. These findings will aid governments and agricultural sectors in making better data-driven decisions for policymakers.

43

The importance of specification of the deterministic components in the co-integration model: using data on employment costs and gross earnings to show the impact of model misspecification

Dr Sagaren Pillay[1]
[1]Statistics South Africa, Pretoria, South Africa

This paper investigates the impact of different specifications of deterministic components in the vector error correction model (VECM) form estimated with Johansen's multivariate maximum likelihood approach. Using time series for employment costs and gross earnings data we show the impact of the misspecification of the deterministic components of the estimated Co-integration model. The study suggests that great care must be exercised in model specification. The inclusion or exclusion of the deterministic trend should be clearly justified to avoid misleading results.

127

Soft clustering missing at random (MAR) data

Mr Jason Pillay[1]
[1]University of Pretoria, Pretoria, South Africa

Analysing high-dimensional data is now standard practice in industry and academia due to constant improvements in computational power as well as the increased availability of high-dimensional data. Model-based clustering is one of the more popular analyses tools because its results are typically easy to interpret. Model-based clustering techniques have a large focus on assuming normality, or equivalently assuming spherical symmetry, as a property of the data. However, real-world datasets frequently include challenges such as asymmetric and leptokurtic clusters. The clustering methods compensate by overestimating the number of clusters than what is present in the data. Furthermore, datasets have rows with missing entries with prominent occurrences in environmental, medical, and economic disciplines. Since many of the current frameworks of model-based clustering assume only complete data at disposal, their usage is severely constrained in applications where partially seen records are typical. Clustering complete data from these partially seen records to apply clustering algorithms reduces statistical power and introduces bias to parameter estimates. In this work, we allow the family of mixtures of scale mixtures of multivariate skew-normal distributions to accommodate missing at random values by using the expectation maximization (EM) algorithm. The algorithm is extended to account for missing values by deriving closed-form expressions that impute the missing values, which are immediately used to fit the model. In this way, we do not disregard the information contained in partially observed data and simultaneously impute the missing values and fit the cluster model.

53

Transferability of GANs-UNet model for informal road detection in underdeveloped areas

Miss Luandrie Potgieter[1], Prof Inger Fabris-Rotelli[1], Dr Renate Thiede[1]
[1]University of Pretoria, Hatfield, Gauteng

Road infrastructure is crucial for economic development and societal well-being in developing countries. Effective road maintenance and monitoring are essential for road safety, reduced travel time, lower vehicular costs, and overall economic growth. Conversely, inadequate monitoring can lead to increased accidents, higher fuel consumption, air pollution, and limited access to markets and services. Enhancing road network monitoring in developing nations is thus vital for sustainable development and improving quality of life. Despite this importance, there is a significant lack of accurate and current road data, especially in countries like South Africa. Existing monitoring programs face challenges such as time-consuming assessments, limited surveillance vehicles, and high costs. Remote sensing (RS) technologies offer a promising solution by providing large-scale, accurate data for national, regional, and informal roads. However, there is no consensus on the best RS algorithms and data sources for road network monitoring, particularly in South Africa. This research aims to develop a road monitoring framework for South Africa, focusing on automated road network extraction, monitoring road quality and condition, and assessing accessibility in townships. By addressing these challenges, this research will contribute to the effective management and maintenance of South Africa's road infrastructure, ultimately benefiting both rural and urban communities and promoting sustainable development.

24

Break detection in high-dimensional panel data

Prof Marie Hušková[1], Prof Charl Pretorius[2]
[1]Charles University, Prague, Czech Republic, [2]Centre for BMI, North-West University, Potchefstroom, South Africa

Panel regression models with cross-sectional dimension N are considered. The aim is to test, based on T observations, whether the intercept in the model remains unchanged throughout the observation period. The test procedure involves the use of a CUSUM-type statistic derived via a quasi-likelihood argument. The limit behaviour under the null distribution of the test under strong mixing and stationarity conditions on the errors and regressors, are presented. Both independent panels as well as the case of strong cross-sectional dependence are considered. A self-normalised version of the test is also proposed, which is convenient from a practical perspective - particularly to avoid estimation of long-run variances. The theoretical results are supported by a simulation study that indicates that the test works well in the case of small to moderate sample sizes. The talk concludes with an illustrative application in the framework of the four-factor CAPM model.

23

GPAbin biplots for continuous data: a methodology for combining biplots of completed continuous data sets

Mr Mokgeseng Ramaisa[1], Dr Johané Nienkemper-Swanepoel[1]
[1]Department of Statistics and Actuarial Science, Centre for Multi-Dimensional Data Visualisation (MuViSU), Stellenbosch University, Stellenbosch, South Africa

Missing data in real word applications are frequently encountered by data practitioners. The strategy to handle missing data is often deletion which results in loss of information and biased results in analyses. It is important to investigate the underlying reason for missingness to decide on an appropriate handling strategy to reduce bias. Multiple imputation is a superior technique to handle missing values, in which multiple possible values for missing data are imputed, resulting in multiple completed data sets. A practitioner may be interested in exploratory data analysis through visualisation, however interpreting and analysing multiple visualisations of completed data sets may lead to subjective bias and may be time intensive. GPAbin has been developed to unify multiple completed multivariate categorical data visualisations, specifically multiple correspondence analysis biplots. This has been achieved using generalised orthogonal Procrustes analysis and Rubin's rules to result in a single unbiased visualisation for practitioners. The extension of the GPAbin methodology to continuous completed data sets is presented in this paper. This is achieved by utilising the extension of the classic Gabriel principal component analysis biplot as a starting point. A simulation study is presented to further understand the performance of the methodology presented.

143

Enhancing research guidance in Statistics supervision: adapting to the generative AI era

Dr Danielle Roberts[1], Prof Inger Fabris-Rotelli[2], Prof Sonali Das[2], Prof Michael von Maltitz[3], Dr Ansie Smit[2], Prof Daniel Maposa[4], Prof Fabio Correa[3]
[1]University of KwaZulu-Natal, Durban, South Africa, [2]University of Pretoria, Pretoria, South Africa, [3]University of the Free State, Bloemfontein, South Africa, [4]University of Limpopo, Polokwane, South Africa

Generative artificial intelligence (GenAI) has the potential to enhance PhD research in Statistics by streamlining complex tasks such as literature reviews, coding, and brainstorming. With GenAI tools assisting in these areas, there is potential for students to accelerate their research progress and focus on more critical aspects of their work. However, the rise of GenAI tools has also introduced challenges, including concerns about over-reliance, ethical implications, and potential misuse, such as students presenting AI-generated content as their own without understanding aspects of the content, thereby compromising academic integrity. This shift requires supervisors to rethink their approach to guiding students, emphasising transparency and the ethical use of AI. As a result, the supervisor-student relationship is evolving, with a growing need for open dialogue on the appropriate use of GenAI tools in research, fostering a balance between innovation, authenticity, and academic rigor.

This study explores the strengths and limitations of using AI tools in general, in Statistics research and supervision, drawing on qualitative findings and sentiment analysis of surveys conducted among Statistics academics and postgraduate students in South African institutions. These explorations and findings have informed the development of a guiding rubric by the authors to provide supervisors and PhD students with practical guidance on responsible use of AI tools during the PhD process. The rubric offers a set of suggested guidelines, outlining best practices for supervision, including the ethical use of AI tools as a complement to independent work.

55

Compositional biplot approaches

Mr Phuti Sebatjane[1,2], Prof Sugnet Lubbe[2], Prof Niël le Roux[2]
[1]Department of Statistics, University of South Africa, Pretoria, South Africa, [2]MuViSU, Department of Statistics and Actuarial Science, Stellenbosch University, Stellenbosch, South Africa

Biplots are powerful visualisation tools for multivariate data. Here we present biplot methodologies for compositional data which are multivariate data with a constant sum constraint and (or) carrying relative meaning. Compositional data are ratio scale in nature and therefore an appropriate transformation needs to be applied to bring the data onto an interval scale for biplot visualisation. Compositional biplots are based on a logratio transformation involving additive logratios, centered logratios or isometric logratios. The focus here is on compositional biplots based on centered logratio coefficients i.e., logratio analysis (LRA) and principal component analysis for compositional data (PCA CoDA). Two biplot methodologies based on LRA and two based on PCA CoDA are compared and the conclusion is that on one hand the methods are essentially similar while on the other hand they are extensions of one another. This conclusion is supported with a data example and matrix results showing the relationship between the methods. The biplot displays are also presented and possible ways to enhance their interpretation are suggested as part of the ongoing work.

146

Bayesian prior elicitation for malaria modelling

Ms Makwelantle Asnath Sehlabana[1], Prof Daniel Maposa[2], Dr Alexender Boateng[3], Prof Sonali Das[4]
[1]University of Limpopo, Polokwane, South Africa, [2]University of Limpopo, Polokwane, South Africa, [3]Department of Mathematics and Computer Science, Modern College of Business and Science, Bawshar, Oman, [4]University of Pretoria, Pretoria, South Africa

Despite the wealth of knowledge in statistical epidemiology, subjective Bayesian methods, which incorporate expert knowledge into disease modelling, remain underutilised. While objective priors are commonly favoured for their simplicity, subjective Bayesian approaches leverage specific prior knowledge, such as expert insights on malaria transmission, leading to more informed models. The aim of this research is to develop a cost-effective approach to improve the elicitation and integration of expert knowledge into Bayesian models for malaria transmission, focusing on the influence of environmental and climatic factors. Prior elicitation, however, poses several challenges. Translating expert judgments into statistical terms such as probability distributions is complex and often requires specialised software. Existing elicitation methods, such as the Sheffield method, are resource-intensive, demanding significant time and cognitive effort from experts and researchers. To address these challenges, this research integrates the Analytic Hierarchy Process (AHP) with statistical validation techniques. Expert knowledge was collected through questionnaires and converted into pairwise comparisons, which were then quantified into AHP weights, representing the importance of various environmental factors in malaria transmission. These weights were fitted to various relevant probability distributions and evaluated using the goodness of fit tests. The results indicated that Gamma and Beta distributions best captured expert knowledge. This approach offers a more practical method for applying subjective Bayesian models in epidemiology by simplifying the elicitation process and reducing the technical burden. Future research will compare these elicited priors with objective priors to assess their impact on model performance across different domains.

105

Comparative analysis of the return level estimates based on block maxima and POT extreme value theory approaches

Ms Anna Seimela[1], Prof Daniel Maposa[1]
[1]University of Limpopo, Polokwane, South Africa

Climate extremes such as floods and heat waves have become serious issues as they are the main causes of natural disasters that affect humans and the environment. These extreme events impact the society and the environment posing serious challenges in developing countries such as South Africa. This study analyses maximum temperatures in the Limpopo province of South Africa through a comparative analysis of the return levels of block maxima and peaks-over-threshold (POT) realisations. The block maxima approach used in the study is the generalised extreme value (GEV) method, while the Poisson point process and generalised Pareto distribution (GPD) was used for the POT approach. The profile likelihood method was used for the 95% confidence intervals. Four stations namely Mara, Messina, Polokwane and Thabazimbi were considered for the study. The findings of the comparative study revealed that the GEV, Poisson and GPD return level estimates were comparable for three stations except for Polokwane where higher return level estimates were observed for the Poisson point process and GPD method as compared to the GEV method. The return level forecasts of 40 ℃ for higher return periods at Thabazimbi suggest that average temperatures of 40 ℃ are expected to be exceeded at least once in 100-years. These findings revealed that the two approaches, block maxima and POT, are comparable for higher return periods. Future studies will explore the inclusion of other methods such as blended GEV and copula methods.

63

Enhancing financial market risk measures: a comparative analysis of long-memory GARCH-type models

Dr Modisane Seitshiro[1]
[1]North-West University, Potchefstroom, South Africa

The purpose of the article is to investigate whether considering stylised facts in financial time series leads to better estimation of risk measures. The study focuses on long-memory GARCH-type models chosen for their ability to capture characteristics of financial time series like long memory and volatility clustering. Heavy-tailed parametric distributions are considered to evaluate the effectiveness of estimating the risk measures. Historical closing prices of two financial market indices are analysed using statistical modelling and rigorous testing. The back testing results showed that the FIGARCH and FIAPARCH models with heavy-tailed distributions were effective in capturing the characteristics of the financial time series, such as heavy tails and volatility clustering, which are essential for accurate risk measurement. The research emphasises that long-memory GARCH-type models generally outperform traditional short-memory GARCH models in estimating extreme risk measures, highlighting their importance in financial risk management. The findings suggest that incorporating long-memory and asymmetric characteristics into risk models can lead to better risk measure estimations, which is crucial for financial institutions in managing market risk effectively. The study contributes to improved risk management practices for financial institutions by demonstrating that long-memory GARCH-type models provide more accurate estimations of risk measures. This is particularly important for ensuring that financial institutions can meet capital requirements and manage market risks effectively. The research sheds light on the stylised facts that are critical for understanding financial market behaviour. This understanding helps market participants make more informed decisions regarding trading strategies and risk assessment. The findings offer practical insights for portfolio managers, risk analysts, and financial regulators by highlighting the importance of using advanced statistical models that account for long memory and asymmetry in financial data.

54

Spatial dependency modelling of disjoint spatial areas - SAPRIN urban node analysis

Ms Ephent Selahle[1], Prof Inger Fabris-Rotelli[1], Mrs Nada Abdelatif[2]
[1]University of Pretoria, Pretoria, South Africa, [2]South African Medical Research Council, Cape Town, South Africa

Spatial data analysis often requires understanding the dependence between polygons, which is crucial for accurate modelling. Spatial weight matrices play a fundamental role in quantifying the spatial and temporal relationships between observations or geographical points. These matrices describe how the value of a variable at one location relates to values at other locations, and their construction differs depending on the research context. In this research, we focus on using Euclidean distance to measure the spatial relationships between units, but we also acknowledge other types of matrices, such as binary contiguity matrices, k-nearest neighbours, and inverse distance weighting. Each matrix type offers unique insights into spatial dependencies. Additionally, detecting spatial autocorrelation, which indicates non-random spatial distributions, is a critical aspect of our analysis. We look more into Moran's I index as a primary tool to assess the degree of clustering or dispersion in spatial patterns.

Beyond spatial dependencies, our research delves into demographic disparities across disjoint urban areas, such as residential, commercial, and mixed-use zones, within a city, based on data collected by SAPRIN. A particular focus is given to how gender influences these patterns. By analysing factors such as age distribution, household size, and ethnic composition, we aim to uncover the relationships between geographic features, urban development distribution, population density, and their correlations with health risk profiles and seropositivity probabilities. These insights are vital for guiding urban planning and public health policies towards creating more equitable and healthy urban environments.

110

Estimating disability rates in South African districts using area-level Poisson mixed models

Prof Yegnanew Shiferaw[1]
[1]Department Of Statistics, University of Johannesburg, Johannesburg, South Africa

The study aimed to estimate disability rates at the district municipality levels in South Africa using small area estimation (SAE) approach. It aimed to address the challenges of producing accurate estimates of disability prevalence in small areas by using two big data sources. Additionally, the study aimed to assess how well these small area estimates aligned with data from national censuses. We utilised SAE models, specifically the area-level Poisson mixed model, to analyse disability status using data from the 2023 Generalised Household Survey (GHS) and the 2022 Population Census. The GHS defines disability status as follows: disability (if an individual has `Some difficulty' or `A lot of difficulty' or is `Unable to do' for one or more categories), UN disability (If an individual has `Some difficulty' for two or more of the six categories), or severe disability (if an individual has `A lot of difficulty' or is `Unable to do' for one or more categories) across the 52 districts in South Africa. The SAE method provides more accurate estimates of disability prevalence at the district level compared to direct estimates based on the 2023 GHS data alone. According to the model, severe disability rates range from DC48 (0.8507%, 95% CI= 0.3326%, 1.4259%) to DC27 (13.4258%, 95% CI=11.6061%, 14.7649%). Meanwhile, model-based disability rates range from DC23 (4.4799%, 95% CI= 2.5521%, 6.4077%) to DC9 (26.3281%, 95% CI=23.9204%, 28.7358%). In addition, model-based UN disability rates range from DC48 (1.7231%, 95% CI= 0.8347%, 2.6199%) to DC27 (14.1957%, 95% CI=12.2681%, 16.3476%). Accurate and insightful estimates of disability rates within small geographical areas, such as district municipalities, can be developed by integrating publicly accessible data sources. This enriched information proves invaluable for local health departments and policymakers engaged in evidence-based decision-making.

25

Data-driven approaches for predicting electricity demand

Prof Caston Sigauke[1]
[1]University of Venda, Thohoyandou, South Africa

In this talk, we discuss short-term forecasting in high-resolution datasets arising from industrial and scientific applications. We will include some basic theories and the practical usefulness of modern statistical methods, especially their functional data analysis features. Although our main case study is about electricity load forecasting, the techniques covered apply to many seasonal problems depending on other exogenous variables. During the last decades, the intersection of statistics and machine learning with software tools has revolutionized how problems in the real world are addressed. The purpose of the talk will be to introduce a set of statistical and machine learning techniques and demonstrate how they can be effectively applied in forecasting, expert systems, and big data analysis.

132

Joint modelling for longitudinal and interval censored survival data

Dr Isaac Luwinga Singini[1], Prof Ding-Geng Chen[2], Associate Prof Freedom Gumedze[3]
[1]Biostatics Research Unit, South African Medical Research Council, Cape Town, South Africa,
[2]Department of Statistics, University of Pretoria, Pretoria, South Africa, Pretoria, South Africa,
[3]Statistical Sciences Department, University of Cape Town, Cape Town, South Africa

Joint models for longitudinal and survival data are a class of models that jointly analyse an out-come repeatedly observed over time such as a bio-marker and associated event times. 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. For the survival sub-model this is done by recording the moments of the event of interest and calculating the time span be-tween the event and some initial onset time. In medical research for instance, interest would be on enrolment into a study and disease progression, which is characterised by HIV positivity and onset of AIDS or using our motivating data set (IMPI trial) it would be characterised by time to constrictive pericarditis (constriction).

The joint modelling framework has mainly focused on right censored data in the survival outcome for the last decade. This has been for two-stage joint model, shared parameter joint models and latent class joint models. There have been many theoretical developments in the last five decades that have focused on censoring mechanisms in order to correctly model time to event data e.g. left or right censoring, however interval censoring has seldom been implemented in the joint modelling framework. This has been due to the fact that many are unaware of the impact of inappropriately dealing with interval censoring within the joint modelling framework. The other complexity has been that the necessary software that handles interval censored data in the joint modelling framework is not readily available. In this chapter we fill the gap between theory and practice by illustrating our theoretical technique using the interval censored data in the joint model using a cardiology multi-centre clinical trial. This is done using R software.

138

An assessment of the impact of spatial connectivity structures on spatial model fit: machine-learning approach

Dr Claris Siyamayambo[1], Dr Edith Phalane[1], Prof Refilwe Phaswana-Mafuya[1], Prof Inger Fabris-Rotelli[2]
[1]University of Johannesburg, Johannesburg, South Africa, [2]University of Pretoria, Pretoria, South Africa

Optimising spatial connectivity structures is paramount in spatial models to help understand the geography that shapes the world. However, there are various spatial connectivity structures available making the selection process of an ideal connectivity structure difficult given diverse spatial relationships and data sets. Moreover, a lack of a set standard to identify the best spatial connectivity structure for a given data set has caused people to make random choices. These arbitrary selections of spatial connectivity structures can negatively impact spatial model fit and predictive accuracy thereby hindering correct decision-making. Spatial models are essential for exploring spatially dependent data; hence, there is a need for appropriate model development using optimal spatial connectivity structures. Machine-learning algorithms can be used to develop a method for the model to choose the best spatial connectivity structure option. This is essential for generating data-driven and automated procedures for determining spatial connectivity structures that maximise model performance. This study will review existing literature and analyse available and popular spatial connectivity structures used in developing spatial models. Real-world spatial datasets will be analysed using supervised machine-learning algorithms to assess the impact of reviewed spatial connectivity structures on spatial model fit. This research advocates for data-driven methods that enhance model performance and make more informed decisions in spatial analysis.

128

Quantifying the directional relationship between natural hydrogen depressions and fault lines in Mpumalanga, South Africa

Mr Calvin Jens Botha[1], Dr Ansie Smit[2], Prof Inger Fabris-Rotelli[1], Dr Najmeh Nakhaei Rad[1], Prof Adam John Bumby[2], Dr Brenda Otieno Mac'Oduol[1]
[1]Department of Statistics, University of Pretoria, Pretoria, South Africa, [2]Department of Geology, University of Pretoria, Pretoria, South Africa

The global energy landscape is currently undergoing several profound changes, developing a path towards accessible renewable resources. Natural hydrogen has emerged as a promising energy source, yet many questions remain about its geological formation. This research explores the directional relationship between seepages of natural hydrogen, known as sub-circular depressions, and geological lineaments in Mpumalanga, South Africa. The methodology employs circular statistics to analyse depression orientations and uses rose diagrams for data visualisation. Comparisons between the angles of depressions and geological lineaments are conducted using an algorithm for data set division, followed by K-means clustering to classify the depressions by size. Robust hypothesis testing, tailored for circular data, is applied through the single-sample likelihood ratio test, along with the Watson-Williams and the Watson-Wheeler tests. Initial findings suggest that the average orientation between the depressions and geological lineaments in the study area differs significantly from zero. Additionally, the Watson-Wheeler test indicates that the three-cluster solutions exhibit no significant directional differences when comparing individual clusters, while the four-cluster solution reveals substantial variation in angular distributions. The findings provide a foundation for understanding these interactions and open the door to exciting possibilities for future research and exploration, potentially revolutionising the way we harness and study the implementation of natural hydrogen as a renewable energy.

129

Profile-likelihood based confidence intervals in earthquake hazard assessment models

Mr Siyamthanda Prusent[1], Dr Ansie Smit[2], Prof Inger Fabris-Rotelli[1], Dr Najmeh Nakhaei Rad[1], Dr Brenda Otieno Mac'Oduol[1]
[1]Department of Statistics, University of Pretoria, [2]Department of Geology, University of Pretoria

Over the past few decades, earthquake hazard assessment distributions have been extensively developed and refined. Many of these distributions, typically non-Gaussian and skewed, account for missing data, uncertainties in earthquake magnitudes (sizes), and uncertainties in the applied distribution. Maximum likelihood estimation (MLE) is commonly used to estimate seismicity parameters, such as the earthquake occurrence rate (λ) and the parameter related to the b-value of the Gutenberg-Richter law (β). However, confidence intervals for these parameters are often symmetric due to the assumption of normality, which may not align with the skewed nature of earthquake magnitude data. The core aim of this research is to derive confidence intervals that reflect this asymmetry by exploring profile likelihood methods. Examples based on synthetic (simulated) earthquake data and actual earthquake data acquired for the Tulbagh area in the Western Cape, South Africa will be presented to determine if the profile-likelihood-based confidence intervals are any different from the Wald-type confidence intervals already presented in theory.

28

A threshold-search approximate Bayesian computation algorithm for parameter estimation

Dr Neill Smit[1]
[1]North-West University, Potchefstroom, South Africa

Approximate Bayesian computation (ABC) represents a class of likelihood-free methods which can be used to approximate posterior distributions. ABC is particularly useful in cases where the likelihood function is difficult to compute analytically or even computationally intractable. In this paper, a simple modification of the standard ABC rejection sampling algorithm is introduced, called the threshold-search ABC algorithm, where the acceptance threshold is adaptively adjusted as candidate points are accepted or rejected. This modification not only enables the acceptance of preferable candidate points but also, to some degree, eliminates the importance of choosing a suitable acceptance threshold. Furthermore, the algorithm can also act as a search mechanism for determining a fixed acceptance threshold for the standard ABC rejection sampling algorithm. A simulation study is conducted to compare the performance of the threshold-search ABC algorithm against maximum likelihood estimation. In the simulation study, parameter estimation for some widely used life distributions is performed, where several distance functions for the threshold-search ABC algorithm are considered.

37

Goodness-of-fit tests with applications in risk modelling

Ms Leoni Snyman[1], Prof James Allison[1], Prof Jaco Visagie[1], Prof Simos Meintanis[2]
[1]North-West University, Potchefstroom, South Africa, [2]University of Athens, Athens, Greece

The empirical Laplace transform is utilised to construct a L2-type test for the null hypothesis that a positive random variable follows a one-sided stable distribution with an unspecified tail-index $\alpha \in (0,1)$. Large-sample properties of the test are investigated. The results of a Monte-Carlo study are presented where the finite sample powers of the newly proposed test is compared to existing tests. Because operation losses are considered to follow a positively skewed and heavy tailed distribution, the positive stable distribution is potentially a good fit to these losses. We thus conclude the talk with the application of the new test to a real-world dataset involving operational risk data.

91

An improved test for the accuracy of spatial point pattern tests

Dr Rene Stander[1], Prof Inger Fabris-Rotelli[1], Prof Gregory Breetzke[1], Dr Jean-Pierre Stander[1]
[1]University of Pretoria, Pretoria, South Africa

In this talk we re-examine the similarity threshold of Andresen's S-index for spatial point patterns. Andresen's S-index is used widely by geographers, specifically in criminology literature, to determine the similarity between two spatial point patterns. A spatial point pattern consists of the locations where an event of interest occurred. The S-index represents the percentage of spatial units that have similar spatial patterns in both point patterns and ranges from 0 to 1. The test is subjective in that it delineates spatial similarity and dissimilarity at a threshold of S = 0.8. We propose a technique to remove this subjectivity by taking into account the second-order nature of the spatial data. An improved, more robust test is thus set up. This approach will be applied to road networks in the city centre of Pretoria and Johannesburg in South Africa. The road network will be represented as a point pattern and the similarity will be determined on whether the same road structures are found.

7

Comparing the asymptotic relative efficiency of the CMP model with the negative binomial model

Dr Yuvraj Sunecher[1]
[1]University of Technology, Pointe Aux Sables, Mauritius

Time series of counts are frequently subject to the dispersion phenomena in real-world situations while also being impacted by certain explanatory variables. The first order integer-valued moving average structure (INMA(1)) with COM-Poisson (CMP) innovations under a link function, is assumed in this paper in order to address these two concerns. The second part of the paper consists of estimating the regression effects, dispersion and the serial parameters using a Generalized Quasi-Likelihood (GQL) approach. In addition, this paper also compares the asymptotic relative efficiency (ARE) of the CMP with the Negative Binomial model in the fitting of MA(1) time series of over-dispersed counts. In the same section, an assessment of the ARE of CMP with Negative Binomial is presented. The conclusion is provided in the last section.

76

A generalised homogeneously weighted moving average scheme for monitoring the process mean

Mr Maonatlala Thanwane[1], Dr Majika Jean Claude Malela[1], Prof Frans Kanfer[1], Prof Kashinath Chatterjee[2]
[1]University of Pretoria, Pretoria, South Africa, [2]University of Augusta, Georgia, United States of America

This paper introduces a generalised version of the homogenously weighted moving average (HWMA) monitoring scheme for monitoring the process mean (HWMA $\bar{X}$). The proposed scheme is constructed using r smoothing parameters (r≥ 1) instead of one as it is the case for the HWMA $\bar{X}$ scheme. Thus, the existing HWMA $\bar{X}$ scheme is a special case of the new generalised HWMA $\bar{X}$ (GHWMA $\bar{X}$) scheme when r= 1. The properties of the new scheme are derived and the in-control (IC) and out-of-control (OOC) performances investigated in terms of the zero- and state-state characteristics of the run-length distribution. In addition, the performance of the GHWMA $\bar{X}$ scheme is compared with that of the cumulative (CUSUM) and exponentially weighted moving average (EWMA) $\bar{X}$ schemes. It is found that the new GHWMA $\bar{X}$ is flexible through the adjustment of the smoothing parameters and outperforms the existing schemes when the weights are well-selected. To demonstrate the application and implementation of the new scheme, numerical examples are provided using real-world and simulated data.

58

A statistical exploration of the effect of road network structure on road-based accessibility

Dr Renate Thiede[1], Prof Inger Fabris-Rotelli[1], Prof Pravesh Debba[2], Prof Christopher Cleghorn[3]
[1]University of Pretoria, Pretoria, South Africa, [2]CSIR, Pretoria, South Africa, [3]University of the Witwatersrand, Johannesburg, South Africa

Accessibility analyses quantify the level of access to certain areas or opportunities, such as employment and healthcare facilities. Since public data is often aggregated at the level of regions, such as administrative units, it is useful to quantify accessibility between regions. Many factors influence inter-regional accessibility, most notably the accessibility metric used, and the way in which regions are chosen. This paper investigates the effects of road network structure on accessibility, using a previously developed inter-regional accessibility model that bases its accessibility metric on travel distance via the road network. This paper considers an area within the City of Tshwane municipality in South Africa. We investigate the effects of road structure in two ways. Firstly, regions are chosen based on the road network structure, which is done by extending a previously developed road network clustering algorithm for this novel use. Different spatial scales of regionalisation are considered, and the accessibility between these regions is compared to the accessibility between administrative units within the study area. Secondly, the effect of road network homogeneity on accessibility is investigated, where homogeneity corresponds to a uniform concentration of roads across a region. The results show that although road network homogeneity does not significantly correlate with accessibility, the way in which regions are chosen, and their spatial scale has a strong effect on the results of the accessibility model. Our novel method of obtaining regions thus provides fresh insights into road-based accessibility within the City of Tshwane.

116

Prevalence and risk factors associated with HIV infection among pregnant antenatal attendees in Limpopo Province

Dr Oratilwe Penwell Mokoena[1], Mr Donald Tshabalala[1], Dr Thembelihle Sam Ntuli[1], Mr IT Boshomane
[1]Sefako Makgatho University

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

51

A quantile regression model for bounded longitudinal data and survival data

Dr Divan A Burger[1,2,3], Dr Sean van der Merwe[2], Dr Janet van Niekerk[4,3], Prof Emmanuel Lesaffre[5], Mr Antoine Pironet[6]
[1]Syneos Health, Bloemfontein, South Africa, [2]University of the Free State, Bloemfontein, South Africa, [3]University of Pretoria, Pretoria, South Africa, [4]King Abdullah University of Science and Technology, Thuwal, Kingdom of Saudi Arabia, [5]KU Leuven, Leuven, Belgium, [6]AARDEX Group, Liège, Belgium

This study introduces a novel joint modelling framework integrating quantile regression for longitudinal continuous proportions data with Cox regression for time-to-event analysis, employing integrated nested Laplace approximation (INLA) for Bayesian inference. Our approach facilitates an examination across the entire distribution of patient health metrics over time, including the occurrence of key health events and their impact on patient outcomes, particularly in the context of medication adherence and persistence. INLA's fast computational speed significantly enhances the efficiency of this process, making the model particularly suitable for applications requiring rapid data analysis and updates. Applying this model to a dataset of patients who underwent treatment with atorvastatin, we demonstrate the significant impact of targeted interventions on improving medication adherence and persistence across various patient subgroups. Furthermore, we have developed a dynamic prediction method within this framework that rapidly estimates persistence probabilities based on the latest medication adherence data, demonstrating INLA's quick updates and prediction capability. The simulation study validates the reliability of our modelling approach, evidenced by minimal bias and appropriate credible interval coverage probabilities across different quantile levels.

166

A noncentral Poison-Lindley distribution contextualised in a process monitoring framework

Dr Ané van der Merwe[1], Prof Johan Ferreira[1]
[1]University of Pretoria, Pretoria, South Africa

Control charts are powerful tools in the context of statistical process control, as they assist in monitoring the behaviour of processes over time. The C-chart plays a vital role in monitoring count data as it checks for the number of nonconformities in equal subgroup sizes. The most common quality characteristic distribution for this chart is the Poisson distribution; however, charting statistics often exhibits overdispersion, and thus the equidispersion assumption of the Poisson distribution no longer holds. A suitable candidate with the ability to capture over- or underdispersion has to be considered to ensure that the likelihood of observing false alarm rates is not inflated. A noncentral Poisson-Lindley distribution is proposed as a suitable candidate in this framework, accounting for additional flexibility in describing and managing overdispersion specifically. This work aims to firstly introduce the distribution with some of its important characteristics. Parametric- and bootstrap control charts are comparatively analysed when the observations are assumed to follow this noncentral Poisson-Lindley distribution. Extensive simulation studies validate this model in this framework, and thoughts for future work are explored.

102

From hotspot detection to accessibility: a spatial network analysis of informal settlements

Mr R van der Walt[1], Dr R.N. Thiede[1], Prof I.N. Fabris-Rotelli[1]
[1]University of Pretoria, Pretoria, South Africa

Accessibility to essential facilities plays a crucial role in our daily lives. Accessibility is defined as the ease with which a facility can be reached. We explore the accessibility of facilities such as police stations, hospitals, and schools from housing locations within the Melusi informal settlement. The rate of houses per road segment is calculated, and statistically significant hot- and coldspots of the rate of houses are identified by employing the Getis-Ord statistic, resulting in areas of high or low spatial clustering. Dijkstra's shortest path algorithm is then applied to assess the accessibility of statistically significant hot- and coldspots to the facilities of interest. A comparison is made between underserved and well-served areas. The results provide insight into the spatial layout of the settlement, and the degree to which residents of the settlement have access to crucial facilities.

45

Mapping soil thickness by accounting for right-censored data with survival probabilities and machine learning

Dr Stephan van der Westhuizen[1,2,3], Prof. G. B. M. Heuvelink[2,3], Dr D. P. Hofmeyr[4], Dr L Poggio[3], Dr M Nussbaum[5], Prof C Brungard[6]
[1]Stellenbosch University, Stellenbosch, South Africa, [2]Wageningen University, Wageningen, the Netherlands, [3]ISRIC - World Soil Information, Wageningen, the Netherlands, [4]Lancaster University, Lancaster, United Kingdom, [5]Utrecht University, Utrecht, the Netherlands, [6]New Mexico State University, Las Cruces, United States of America

In digital soil mapping, modelling soil thickness poses a challenge due to the prevalent issue of right-censored data. This means that the true soil thickness exceeds the depth of sampling, and neglecting to account for the censored nature of the data can lead to poor model performance and underestimation of the true soil thickness. Survival analysis is a well-established domain of statistical modelling that can deal with censored data. The random survival forest is a notable example of a survival-related machine learning approach used to address right-censored soil property data in digital soil mapping. Previous studies that employed this model either focused on mapping the probability of soil thickness exceeding certain depths, and thereby not mapping soil thickness itself, or dismissed it due to perceived poor performance. In this study we propose an alternative survival model to map soil thickness that is based on the inverse probability of censoring weighting. In this approach, calibration data are weighted by the inverse of the probability that soil thickness exceeds a certain depth, that is, a survival probability. These weights can then be used with most machine learning models. We used the weights with a regular random forest, and compared it to a random survival forest, and other strategies for handling right-censored data, through a comprehensive synthetic simulation study and two real-world case studies. The results suggest that the weighted random forest model produces competitive predictions, establishing it as a viable option for mapping right-censored soil property data

169

An adjustive rating system for rugby union based on exponential smoothing

Dr Paul J van Staden[1], Mr Waldo Botha[1], Mr Carlo Geyer[1]
[1]Department of Statistics, University of Pretoria, Pretoria, South Africa

World Rugby's probit rating model for national teams in rugby union is an adjustive rating system in which the two competing teams in a match exchange rating points based on a comparison between the match result and the predicted match outcome. For the match result, the two teams are assigned 1 for a win and 0 for a loss respectively, while both teams are assigned 0.5 when the match result is a draw. The predicted match outcome, also scaled from 0 to 1, is the probability of a team beating the opponent and is calculated in terms of the relative strength of each team and, if applicable, home advantage. With World Rugby's model, the match result is considered more important than the margin of victory. Consequently, a team does not gain any rating points for a narrow loss to an opponent with a higher rating. Therefore, in this research an adjustive rating system based on exponential smoothing is presented in which margin of victory is directly used in the calculation of the teams' rating points. The two rating systems are applied to and compared for the 16 teams in the United Rugby Championship (URC).

113

Bayesian variable selection for skew-normal models

Mr Arnold van Wyk[1], Prof Andriette Bekker[1], Prof Mohammad Arashi[2], Dr Janet van Niekerk[1,2]
[1]University of Pretoria, Pretoria, South Africa, [2]Ferdowsi University of Mashhad, Mashhad, Iran, [3]King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

Variable selection is one of the most commonly faced problems in statistical analysis. In the frequentist paradigm, penalized regression methods such as L1 regularization and LASSO are used to induce sparsity in high-dimensional settings. In the Bayesian setting, sparsity is typically induced by means of a two-component mixture prior with sufficient probability mass at zero. There has also been a recent development that uses global-local shrinkage priors for high-dimensional Bayesian variable selection. The Dirichlet-Laplace (DL) prior is a popular example of this and has shown promising results compared to existing feature selection methods in the Bayesian framework. In this paper, we propose incorporating an asymmetrical component into the variable selection framework. This is showcased by incorporating a skew-normal random error component into the Dirichlet-Laplace prior for linear regression. We also propose a framework for prior selection and hyperparameter tuning of the proposed model. The performance of the proposed model is assessed and compared with its symmetrical counterpart in both simulated and real-data examples, and is found to not only perform well, but is also able to identify certain non-zero signals due to the inclusion of skewness in the proposed model.

61

Insights into the construction of alternative bivariate cardioid distributions

Mrs Delene van Wyk-de Ridder[2], Prof Johan Ferreira[1], Prof Andriëtte Bekker[1]
[1]University of Pretoria, Pretoria, South Africa, [2]University of Cape Town, Cape Town, South Africa

Bivariate circular data arise in many different disciplines and the development and application of appropriate parametric approaches play an important role in deriving meaningful results of such data. Bivariate circular distributions have been considered for modelling torsional angles in molecules, structural protein bioinformatics, wind direction, and the orientation of scrub bird nests. Bivariate generalisations of the well-known cardioid distribution have been investigated in the literature by relying on a mixture approach. Specifically, marginal (univariate) cardioid distributions are blended together based on the concentration parameter assumed to be beta distributed. These types of mixture models allow the joint probability density functions to be represented as Fourier series, aiding in parameter estimation. The mixture methods are advantageous in the sense that samples can easily be generated for Monte Carlo simulation studies and have straightforward extensions to multivariate cases. In this work, we investigate alternative choices for the distribution of the concentration parameter, such as the non-central beta and non-central bivariate beta distributions, as meaningful parametric alternatives to enrich the methodology in constructing bivariate circular models. In addition, we also investigate the use of a newly proposed bivariate circular model as opposed to the bivariate cardioid. The resultant effect of the parameters from the linear mixing beta choices on the proposed bivariate circular models is of particular interest as we investigate probabilistic behaviours on different manifolds. Some key theoretical aspects are considered and are accompanied by numerical illustrations and results.

100

The road not taken: spatial network optimisation on South African informal settlements

Ms C van Zyl[1], Dr R.N. Thiede[1], Prof I.N. Fabris-Rotelli[1]
[1]University of Pretoria, Pretoria, South Africa

An informal road network is a system of roads which develop without formal planning or design. These networks can be modelled as spatial linear networks. Wherever spatial linear networks are applied, optimisations to that network can also be applied. These optimisations are particularly relevant since there are inherent costs to travelling on a network, such as distance of the road segments, and topographical changes. This research considers the cost of travelling the road segments between the recorded dwellings and other points of interest, such as hospitals, police stations, and schools. The research explored the optimisation of an informal settlement in Gauteng, South Africa. Different existing shortest path algorithms, namely Dijkstra, Bellman-Ford and Floyd-Warshall, were compared. A custom routing algorithm was proposed in which topography was considered in addition to distance when calculating shortest paths on the network, by modifying the existing Dijkstra's algorithm. Finally, the routing obtained between dwellings and points of interest were compared between shortest path functions considering only distance, and the proposed custom routing algorithm considering both distance and topography.

83

Sample size calculations in diagnostic accuracy studies with frequentist and Bayesian approaches

Mrs Lizelle Venter[1], Prof Ding-Geng Chen[2], Prof Inger Fabris-Rotelli[1]
[1]University of Pretoria, Pretoria, South Africa, [2]Arizona State University, Phoenix, U.S.A.

In the age of AI-driven diagnostics, there is a growing need for reliable and flexible methods to estimate sample sizes in diagnostic accuracy studies. These studies evaluate a test's ability to accurately identify the presence or absence of a medical condition. The credibility of such studies depends on the design phase, where calculating the sample size is critical. Incorrectly estimated sample sizes can lead to wasted resources if overestimated or inconclusive study results if underestimated. This paper thoroughly reviews sample size determination methods in diagnostic accuracy studies, comparing frequentist and Bayesian approaches and highlighting their respective strengths and limitations. Additionally, this review explores the application of these methods to various primary endpoints, including sensitivity, specificity, and area under the receiver operating characteristic curve.

114

Construction and analyses of complete diallel cross through partially balanced incomplete block designs

Mr Anteneh Yalew[1], Prof M.K. Sharma[2]
[1]University of the Witwatersrand, Johannesburg, South Africa, [2]Addis Ababa University, Addis Ababa, Ethiopia

Experimental designs for complete diallel cross system IV with equal number of replications of each cross has been studied extensively in the literature. In this paper, we proposed incomplete block designs for complete diallel cross system IV with unequal number of replications for crosses through two-associate classes of partially balanced incomplete block l designs where none of two treatments which are ith associates occur together in exactly blocks (i =1 ,2) is zero. In addition to the block effects and general combining abilities effects, the model also includes the parameter of specific combining abilities effects. The procedure of analyses of these designs has also been developed. The analysis includes the analysis of variance and the estimation for intra-block, inter-block and combined of general and specific combining abilities. Tests of certain hypotheses concerning some general parameters are also given. The optimality of the proposed designs is also considered in comparison to randomized complete block designs. The method of analysis of intra-block, inter-block and combined has been illustrated with the help of numerical data.

99

Geospatial small area estimation of hemoglobin levels of women and children in official statistics

Dr Seyifemickael Amare Yilema[1,2], Dr Najmeh Nakhaei Rad[1], Prof Ding-Geng Chen[1,3]
[1]Department of Statistics, University of Pretoria, Pretoria, South Africa, [2]Debre Tabor University, Debre Tabor, Ethiopia, [3]College of Health Solution, Arizona State University, Phoenix, USA

Hemoglobin, an iron-rich protein in red blood cells, enables the delivery of oxygen to all body tissues. Health problems can arise when hemoglobin levels are either too high or too low, based on specific threshold values. Low hemoglobin levels are commonly associated with anemia, while elevated levels may indicate serious underlying health conditions. Children and women are among the most vulnerable by the low levels of hemoglobin particularly in low- and middle-income countries. Surveys are often providing reliable estimates of target variables for the population at both national and regional levels. However, estimates are unreliable for local areas (zones in our case) in Ethiopia since they are not planned domains in survey designs and have small sample sizes. On the other hand, small-area estimation borrowed strength from neighbouring areas and census data to provide a reliable estimate of the local zones. Therefore, our prime objective is to provide reliable and precise estimates of hemoglobin levels for women and children using geospatial small area estimation by integrating the survey and census datasets. The global and local indicators of spatial association (LISA) statistics for Moran's I and Geary's C tests indicate the existence of spatial autocorrelation among nearby locations in Ethiopian zones. The geospatial small area estimates on the Ethiopian zonal maps show that there are spatial disparities in hemoglobin levels for both women and children throughout the Ethiopian administrative zones. Geospatial small area estimates under the Fay-Herriot model provide relatively more precise and reliable estimates than both the direct and traditional small area estimates across zonal levels of Ethiopia. Policymakers will significantly benefit from these local area-level estimates, and researchers will gain methodological insights into the application of geospatial small-area estimation.

26

Application of longitudinal multilevel zero-inflated Poisson regression in modelling infectious diseases among infants in Ethiopia

Mrs Bezalem Eshetu Yirdaw[1], Prof Legesse Kassa Debusho[1], Dr Aregash Samuel[2]
[1]University of South Africa, Johannesburg, South Africa, [2]Ethiopian Public Health Institute, Gulele Sub City, Ethiopia

In sub-Saharan African countries, preventable and manageable diseases such as diarrhea and acute respiratory infections still claim the lives of children. Hence, this study aims to estimate the rate of change in the log expected number of days a child suffers from diarrhea (NOD) and flu (NOF) among children aged 6 to 11 months at the baseline of the study. This study used secondary data which exhibit a longitudinal and multilevel structure. Based on the results of exploratory analysis, a multilevel zero-inflated Poisson regression model with a rate of change in the log expected NOD and NOF described by a quadratic trend was proposed to efficiently analyse both outcomes accounting for correlation between observations and individuals through random effects. Furthermore, residual plots were used to assess the goodness of fit of the model. Considering subject and cluster-specific random effects, the results revealed a quadratic trend in the rate of change of the log expected NOD. Initially, low dose iron Micronutrient Powder (MNP) users exhibited a higher rate of change compared to non-users, but this trend reversed over time. Similarly, the log expected NOF decreased for children who used MNP and exclusively breastfed for six months, in comparison to their counterparts. In addition, the odds of not having flu decreased with each two-week increment for MNP users, as compared to non-MNP users. Furthermore, an increase in NOD resulted in an increase in the log expected NOF. Region and exclusive breastfeeding also have a significant relationship with both NOD and NOF. The findings of this study underscore the importance of commencing analysis of data generated from a study with exploratory analysis. The study highlights the critical role of promoting EBF for the first six months and supporting children with additional food after six months to reduce the burden of infectious diseases.

41

The future of programming in the age of GenAI

Mr Andre Zitzke[1]
[1]SAS, Johannesburg, South Africa

Generative Artificial Intelligence has significantly impacted every aspect of our lives since its introduction in late 2023. Many people are concerned that AI will eventually take over their jobs, and the field of data science is no exception. Programming is a crucial part of a data scientist's daily activities, leading to widespread debate about whether generative AI will replace traditional programming as we know it. This paper summarises the perspectives of various industry leaders on the future of programming in the era of generative AI.