SASA2024 Poster Abstracts

Table of Contents

159 Study of risk factors associated with hypertension: a case study of Dikgale Village, Limpopo Province
Mr Happy Maluleke[1]
[1]University of Limpopo

33 Fault detection and diagnostic analysis in multivariate compositional data
Mr Mduduzi Maphosa[1], Prof Roelof Coetzer[1], Dr Shawn Liebenberg[1], Prof Charl Pretorius[1]
[1]North-West University, Potchefstroom, South Africa

13 Exploring the dynamics of the ZAR/USD exchange rate volatility using FGARCH and First-Order Beta-Skew-T-EGARCH models
Mr Dzulani Mashavhela[1], Dr Thakhani Ravele[1], Prof Caston Segauke[1]
[1]University of Venda, Thohoyandou, South Africa

112 Changes on students preferred learning style over a period of three years (2021 – 2023)
Mr Gezani Richman Miyambu[1], Dr Tshepo Ramarumo[1]
[1]Sefako Makgatho Health Sciences University

86 A Bayesian statistical evaluation of the competition indices used in eucalyptus tree growth modelling
Mr Samuel Mnisi[1], Dr Johannes Hugo[2], Dr Khehla Daniel Moloi[1]
[1]University of Limpopo, Polokwane, South Africa, [2]Nelson Mandela University, Gqeberha, South Africa

119 Survival analysis of patients with hypernephroma
Ms Mamelang Molaba[1], Mr Gezani Richman Miyambu[1]
[1]Sefako Makgatho Health Sciences University, Pretoria, South Africa

56 A shared frailty model for left-truncated and right-censored under-five child mortality data in South Africa
Dr Tshilidzi Benedicta Mulaudzi[1]
[1]University of Venda, Thohoyandou, South Africa

68 Impact of college location on learner's mathematics performance in Limpopo: a correspondence analysis approach
Mr Roland Fomum Nde[1], Prof John Olaomi[1], Prof Legesse Kassa Deboshu[1]
[1]University of South Africa, South Africa

92 Recent advances in spatial statistics methods for rail networks
Ms Nomly Ngubeni[1], Prof Inger Fabris-Rotelli[2], Dr Najmeh Nakhaei Rad[2]
[1]University of Pretoria / Transnet, Pretoria, South Africa, [2]University of Pretoria, Pretoria, South Africa

153 On classes of consistent tests for the Type I Pareto distribution based on a characterisation involving order statistics
Dr Thobeka Nombebe[1], Prof James Allison[1], Prof Leonard Santana[1], Prof Joseph Ngatchou-Wandji[2]
[1]North-West University, Potchefstroom, South Africa, [2]Universite' de Rennes (EHESP) & Institut Elie Cartan de Lorraine, Nancy, France

175 Multinomial regression models: an applied approach to model consumer utility and preference
Mr Macdonald Phasha[1], Dr Judy Kleyn[1]
[1]University of Pretoria, Pretoria, South Africa

47 A comparison of the robust zero-inflated and hurdle models with an application to maternal mortality
Mr Phelo Pitsha[1], Mr R Chiruka[1], Dr C Marange[1], Dr M Mutambayi[1]
[1] University of Fort Hare, Alice, South Africa

117 The prevalence and spatial dynamics of housebreaking and home robbery hotspots in South Africa
Mr Gomolemo Rakale[1], Ms Moleseng CM Ramaube, Mr Sonnyboy K Manthata, Prof Solly M Seeletse
[1]Sefako Makgatho Health Sciences University, Ga-Rankuwa, South Africa

157 Comparison of the discrete-time survival model and machine learning models
Mr Audrey Tshepho Ramachela[1]
[1]University of Venda, Moletjie, South Africa

115 Work integrated learning challenges in a specific academic department of a Gauteng-based University
Dr Tshepo Ramarumo[1], Mr Gezani Richman Miyambu[1]
[1]Sefako Makgatho Health Sciences University, Ga-Rankuwa, South Africa

11 Predicting the closing price of cryptocurrency Ethereum
Dr Thakhani Ravele[1], Mr Vhukhudo Ronny Rambevha[1], Prof Caston Sigauke[1]
[1]University of Venda, Thohoyandou, South Africa

8 A new alpha power Weibull model for analysing time-to-event data: application to diabetes mellitus data
Assistant Prof Getachew Tekle[1], Dr Gao Shengjie, Dr Alisa Craig
[1]Wachemo University, Hossana, Ethiopia

145 Multivariate techniques application to reveal mutual trends among data sources: a consumer research case study
Mrs Marieta van der Rijst[1], Ms Marbi Schwartz[2], Dr Jeannine Marais[2]
[1]Agricultural Research Council, Stellenbosch, South Africa, [2]Stellenbosch University, Stellenbosch, South Africa

149 Modelling the probability of default using logistic regression and threshold-logistic regression
Miss Monalisa Williams[1], Miss Lesego Sepato[1]
[1]Nelson Mandela University, Port Elizabeth, South Africa

131 Modelling and forecasting headline inflation in Ethiopia by supervised machine learning approach

Mr Teklu Nega Yimenu[1], Assistant Prof Dereje Danbe[2], Dr Zeyitu Asfaw[3]
[1]Oda Bultum University, Chiro, Ethiopia, [2]Hawassa University, Hawassa, Ethiopia, [3]Addis Ababa University, Addis Ababa, Ethiopia

96

Application of mixture Weibull-generalized Pareto distribution

Mrs Tolulope Adeniji[1], Dr Akinwumi Odeyemi[1], Dr Chioneso Marange[1]
[1]University of Fort Hare, Alice, South Africa

Weibull and Pareto distributions are widely used in several areas including lifetime data modelling and reliability analysis. In real-life practice, these distributions may not capture the various distributional properties of certain datasets. The use of finite mixture models has enabled it to capture complex patterns in data that a single parametric model may not be able to detect, making it highly adaptable and accurate. This adaptability has resulted in successful application of finite mixture models across different fields. Of particular interest, the Mixture Weibull Pareto (IV) distribution has been used in modelling insurance claims; compared with other distributions and showed superior performance. The current study applies Mixture Weibull Pareto (IV) distribution to health and environmental datasets, which contain real-life events. The performance of Mixture Weibull Pareto (IV) distribution was compared with other distribution models, adopting some selected information criteria, such as Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and Kolmogorov-Smirnov test (KS test) for the significance of the distribution. The results show that Mixture Weibull Pareto (IV) distribution outperformed other distributions, in terms of selected criteria.

162

Autonomous anomaly detection of orchard tree crown delineations

Mr Edward Baleni[1]
[1]University of Cape Town, Cape Town, South Africa

Precision agriculture is a science gaining traction in orchard management studies. It is an advanced monitoring methodology that uses sensors and computer analysis to aid farmers in improving crop yields, utilisation of resource inputs, and management strategies. Sensors are fitted to unmanned aerial vehicles to capture aerial photogrammetry images of orchards at different wavelengths. Serving as a rich data source that can be analysed using computer vision techniques, precision agriculture in orchard management is highly dependent on the high-throughput phenotyping of tree crowns to collect decision-making information. Consequently, this suggests that the most critical component of the analysis is the extraction or delineation of the "individual" tree crowns. This delineation defines the boundary and location where a single tree appears to exist. The delineation of each tree can be extracted using individual tree crown detection and delineation algorithms. Over time, these algorithms have evolved from classical machine vision techniques to more reliable, automatic and generalisable machine learning techniques. While these methods effectively define the boundary of the tree, they are not without error. Poorly estimated delineations limit the potential for automated individual tree crown detection and delineation in orchard management studies. These poor delineations appear as false positives, under-segmented and over-segmented delineations. There is a need to identify these irregular patterns for correction. This work explored several automatic and generalisable anomaly detection techniques to identify these outlying tree crown delineations. This study utilised many popular outlier detection techniques. Local outlier detection methods could detect local and global outliers while maintaining better performance accuracy than global methods. A novel approach, utilising the Local Geary C statistic in a multivariate context, was also tested as a local outlier technique. This demonstrated adequate performance.

A Bayesian statistical evaluation of the competition indices used in eucalyptus tree growth modelling

Mr Samuel Mnisi[1], Dr Johannes Hugo[2], Dr Khehla Daniel Moloi[1]
[1]University of Limpopo, Polokwane, South Africa, [2]Nelson Mandela University, Gqeberha, South Africa

Eucalyptus is the most often planted genus tree in the world which is grown in more than one hundred different countries. It is a subfamily of the Myrtaceae family. The aim of the study was to evaluate individual tree growth as well as the competition present in eucalyptus plantations measured as a function of the growth rate during a particular growth period. A Bayesian re-purposed mixed effect variance component model, known as the Sire-Model in animal breeding problems was used to find the marginal posterior distributions of the unknown parameters and provided the estimates. Estimated tree values were used to make probabilistic statements to rank the places of the trees. The marginal posterior densities were observed using the Gibbs sampler. An identity matrix was used in Gibbs sampling when it was assumed that there is no competition between the trees. A distance independent competition index called the Lorimer (1983), was used to generate a matrix that was used in Gibbs sampling when it was assumed that there is competition between the trees. Results assuming no competition, revealed that the estimated marginal posterior densities of the error variance and tree variance, were slightly positively skewed and severely positively skewed, respectively, like the case when no competition was assumed. However, when competition between the trees was considered, the growth indices for tree 182, tree 184, tree 214 and tree 216 appeared to be lower than the case when no competition was assumed. The overall results of this study obtained were based on a distance-independent competition index, it is advised that an investigation be conducted using a distance-dependent index as well.

Markov-switching volatility models with heavy-tailed distributions for COVID-19 death cases in South Africa

Mr Thembhani Chavalala[1], Prof Retius Chifurira[2], Prof Knowledge Chimhamu[3]
[1]University of Limpopo, Wetton, South Africa, [2]University of KwaZulu-Natal, Wetton, South Africa, [3]University of KwaZulu-Natal, Durban, South Africa

To date, the world has paid a high toll in terms of human lives lost, economic repercussions, and increased poverty due to the coronavirus disease 19 (COVID-19) pandemic. The outbreak of COVID-19 worsened socio-economic conditions and livelihood insecurity for marginalised communities across the globe. South Africa was the only African country in the top 50 countries with the highest number of COVID-19 cases. Hence, in this study, we explore the Markov switching volatility models with heavy-tailed distributions using the daily number of death cases due to COVID-19 dataset in South Africa and three of its nine provinces (Gauteng, KwaZulu-Natal, and Western cape). These three provinces were considered the COVID-19 epicenters in South Africa. The heavy-tailed distributions considered in this study are student-t (StD), skewed student-t (SStD), generalised hyperbolic (GHD), Pearson type IV (PIVD), and generalised lambda distribution (GLD). The sign bias test was conducted to investigate the existence of the asymmetric effects in the returns of the data. Anderson-Darling test was utilised to assess the fitness of the fitted models. The reliability and accuracy of the Value-at-Risk (VaR) of the death due to COVID-19 were assessed using the Kupiec back-testing procedure. The various techniques used depicts that the Markov-switching generalised autoregressive conditional heteroscedasticity (MS-GARCH) model combined with all the studied heavy-tailed distributions fit the return of the data well except for Gauteng where the exponential MS-GARCH (MS-EGARCH) was fitted to capture the asymmetric effects. Moreover, SStD outperforms the other distributions in almost all the considered VaR levels, except for Western cape where some of the VaR estimates are rejected. PIVD and GLD performs better in Western cape at 90% and 97.5% VaR levels, respectively.

111

Progress on the national sunflower-, soybean- and maize cultivar recommendations in South Africa

Mrs Nicolene Cochrane[1], Mrs Annelie de Beer[2], Mrs Safiah Ma'ali[2], Mrs Elmarie Swiegelaar[3]
[1]ARC-CO; Agrimetrics, Pretoria, South Africa, [2]ARC-Grain Crops, Potchefstroom, South Africa, [3]ARC-CO, Pretoria, South Africa

The refining of the national sunflower-, soybean- and maize cultivar recommendations are very important for ensuring food security in South Africa and the world, and optimum profit for farmers in very difficult economic times / circumstances.

The importance of the correct data collection and analyses cannot be emphasized enough; to draw the right conclusions and best recommendations; for the three different crop types (sunflower, soybeans, and maize); for the different climate regions, as well as the different year circumstances, respectively.

For each crop, combined analysis of variance (ANOVA) over the localities were done. Thereafter regression analyses for each cultivar (dependent variable) and locality means (independent variable) were done in order to get the yield probability tables at a specified yield potential, derived from the normal distribution. T- groupings were also calculated, to do the different recommendations, especially where insufficient localities per crop, climate region and year were established.

An application is planned for development for the different crops per climate region over years to refine and to simplify recommendations specifically for the farmer, researchers, and seed companies.

174

Automated analysis of penguin-borne videos using deep learning

Mr Tagen de Wet[1]
[1]University of Cape Town, Rondebosch, South Africa

This study explores the application of deep learning to automate the analysis of video data obtained from animal-borne loggers attached to Chinstrap Penguins (Pygoscelis Antarctica) foraging in Antarctic waters. These video loggers provide a unique "penguin's-eye view" of their behaviours within their natural environment, which normally would not be possible as they are free-ranging and usually forage far from direct observation. Monitoring these behaviours is essential for answering key ecological questions and supporting conservation efforts, as penguin activity can serve as an indicator of ecosystem health. However, the major challenge lies in the sheer volume of raw data collected, which makes manual analysis time-consuming and inefficient. This often results in large datasets being underutilized, limiting their potential to inform ecological research.

Our project addresses this bottleneck by developing deep learning models to automate the extraction of critical information from these videos, streamlining and reducing the manual effort usually required in the ecological data analysis process. Our approach leverages state-of-the-art image classification and object detection techniques to automatically obtain information about the environment such as the number of penguins present in each frame and if a frame contains prey, as well as models suited to classify penguin behaviours such as feeding events.

By automating these tasks, we enhance the efficiency of ecological monitoring, enabling researchers to better understand penguin foraging patterns and their responses to environmental changes.

68

Impact of college location on learner's mathematics performance in Limpopo: a correspondence analysis approach

Mr Roland Fomum Nde[1], Prof John Olaomi[1], Prof Legesse Kassa Deboshu[1]
[1]Unisa, South Africa

This study employs the Correspondence Analysis (CA) technique in order to examine the impact of college location on the mathematics performance of level 4 students in South Africa's Limpopo province. Level 4 students are the final year students in the National Certificates Vocational (NCV) program at Technical and Vocational Education and Training (TVET) colleges. The Chi Square test and F-test (in ANOVA) were conducted on the final year mark of students from three different districts/locations (i.e., rural, urban, and semi-urban areas) in the province. According to the results from the study, mathematics remains a huge problem in the province as less than 50% of the students are passing the subject each year irrespective of their locations. However, students from urban areas are consistently performing better in the subject than their peers from semi-urban and rural areas.

72

Comparison of malaria prevalence among children under five years of age in Mali and Nigeria

Miss Zama Khumalo[1], Prof Sileshi Melesse[1], Prof Henry Mwambi[1]
[1]University of KwaZulu-Natal, Pietermaritzburg, South Africa

INTRODUCTION
Malaria continues to be a severe public health concern in Sub-Saharan Africa, particularly in children under the age of five. This study compares malaria prevalence among children under five in Mali and Nigeria, two nations significantly affected by the illness. The primary goal of this study, which used data from both nations' 2021 Demographic and Health Surveys (DHS) and Malaria Indicator Surveys (MIS), was to determine socioeconomic, environmental, and demographic variables related to childhood malaria infection.

METHODS & RESULTS
Children's malaria status was identified as positive or negative using rapid diagnostic tests (RDTs), and logistic and survey logistic regression models were used to analyse the data. Key determinant factors were the wealth index, home type, availability of clean water, and maternal education level. The findings revealed considerable disparities in malaria risk variables between the two nations. Children aged 48-59 months had the highest malaria risk in both Mali and Nigeria, with older children being more vulnerable. In Mali, children from families with dirt or sand floors or without suitable roof materials were more likely to get malaria, but these characteristics were less significant in Nigeria.

CONCLUSION
The study identified various socio-economic and environmental factors that influence malaria risk in Mali and Nigeria. There are differences in the specific variables that significantly impact malaria risk between the two countries. To raise awareness about malaria prevention, focus on public health campaigns focusing on the specific risk factors identified in each country. Community health workers can consider this information and preventative tools like water treatment solutions and nets.

60

Development of robust imputation techniques with a view to applications in machine learning

Ms Maria Lekganyane[1], Prof Temesgen Zewotir[2], Dr Ahmed Audu[3,4]
[1]Department of Statistical Science, Sefako Makgato Health Science University, Pretoria, South Africa, [2]School of Mathematics, Computer Science and Statistics, University of Kwazulu-Natal, Durban Westville, South Africa, [3]Department of Mathematics and Applied Mathematics, Sefako Makgato Health Science University, Pretoria, South Africa, [4]Department of Statistics, Usmanu Danfodiyo University, Sokoto, Nigeria

In this study, we proposed an imputation method for estimating missing observations or non-response based on a calibration approach, with a view to applications in machine learning. The estimators of the schemes as well as their statistical properties were derived. The efficiency and predictive ability of the schemes were empirically analysed in comparison with some existing imputation schemes. The results of the empirical studies revealed that the proposed imputation schemes are more robust and efficient.

A statistical analysis of factors associated with hypertension among elderly persons in South Africa

Ms Kgethego. Sharina Makgolane[1], Mr Tshepo Frans Maja[1], Ms Nombulelo Porcentia Shongwe[1], Prof Daniel Maposa[1]
[1]University of Limpopo, Polokwane, South Africa

Hypertension is a serious condition in adults, affecting nearly one billion people globally and accounts for at least 15 million lives annually. It is also a common age-related disorder that can lead to cardiovascular and kidney complications. This study is intended to identify key factors to improve policies and prevent hypertension among elderly persons. Utilizing a secondary data from Statistics South Africa website, of 2021 South African Demographics and Health Survey. The survey contains a wealth of information on the individual characteristics of 1353 elderly South African persons aged 60 years and older, focusing on categorical demographic and health factors. Univariate and multivariate binary logistic regression analyses were utilised to determine the factors linked to hypertension in the elderly population. The results of the descriptive analysis indicates that the Eastern Cape region had the highest participation (24.6%) and a 65.5% hypertension rate, followed by KwaZulu-Natal with 21.3% and a 59.7% hypertension rate. Other regions had 5-12% participation, with hypertension rates ranging from 44.4% to 71.7%. Gender-wise, 61.2% were female, with a 68.1% hypertension rate, while males constituted 33.6%, with a 51.5% hypertension rate. The African/Black racial group constituted the majority (79.8%), with a 63.5% hypertension rate, and other racial groups showed varying participation and hypertension rates. Additionally, the multivariate analysis in this study revealed that province of residence, gender, educational level, diabetes, tuberculosis and exercising were found to be significantly associated with hypertension among elderly persons in South Africa. The findings highlight the vulnerability of specific demographic groups, particularly those residing in North West, Mpumalanga, Northern Cape, Free State, Western Cape, and Eastern Cape, as well as individuals with no formal education and those with diabetes, to hypertension. Recognising these associations is crucial for developing targeted interventions and improving healthcare strategies to mitigate hypertension's impact on South Africa's elderly population.

Study of risk factors associated with hypertension: a case study of Dikgale Village, Limpopo Province

Mr Happy Maluleke[1]
[1]University of Limpopo

Hypertension has been shown to be one of the leading public health threats worldwide, mainly due its high frequency and concomitant risks of cardiovascular, stroke and kidney diseases. There is a need for identification of risk factors in a specific area for better and more targeted prevention and control of hypertension and its consequences. The objective of this study was to determine the prevalence of hypertension and its associated risk factors in the Dikgale region of the Limpopo province. The study utilised secondary data collected from the Dikgale village. We present the results gained from the binary logistic regression. Backward elimination method was applied to eliminate variables that were not significant. The study consisted of 11315 people (3985 males and 7330 females) living within the Dikgale region. Descriptive statistics revealed that 1489 (13.2%) out of 11315 people were diagnosed with hypertension. Hypertension was mostly found on females between the age-group of 45-54 to 75-84 years compared to males where the highest was recorded being the between the age-group of 65-74 years. The model considered six factors (age, gender, body mass index (BMI), diastolic, systolic, and hospital visit), five of which were found to have significantly influenced the performance of human blood pressure (hypertension) at 5% level of significance. Only diastolic was insignificant with p-value of 0.0804. The study recommends that conscious efforts be made, and time set aside to provide health education to the communities within the Dikgale village about the risks posed by hypertension.

33

Fault detection and diagnostic analysis in multivariate compositional data

Mr Mduduzi Maphosa[1], Prof Roelof Coetzer[1], Dr Shawn Liebenberg[1], Prof Charl Pretorius[1]
[1]North-West University, Potchefstroom, South Africa

Data-driven fault detection procedures permit the identification of anomalous behaviour, process interruptions, and system breakdowns within industrial facilities. In commercial chemical processes, feed and product streams are often compositional in nature. Compositional data, characterised by constrained non-negative proportions and interdependencies among components, pose unique challenges that necessitate the development of specialised methodologies. This poster presents a statistical framework to detect process faults in multivariate compositional data retrospectively. The data is assumed to follow the Dirichlet distribution, and a likelihood ratio test is developed, together with the associated asymptotic critical values, to detect faults in compositional data. It is shown that the test has the power to detect a single fault across different sample sizes and dimensions.

Exploring the dynamics of the ZAR/USD exchange rate volatility using FGARCH and First-Order Beta-Skew-T-EGARCH models

Mr Dzulani Mashavhela[1], Dr Thakhani Ravele[1], Prof Caston Segauke[1]
[1]University of Venda, Thohoyandou, South Africa

The impact of exchange rate volatility on international trade, investment decisions and economic stability has been a matter of substantial interest to economists, policy-makers and market participants for an extended period. This study investigates the dynamics of the ZAR/USD exchange rate volatility using advanced econometric models: Family FGARCH (FGARCH) model and the First-Order Beta-Skew-T-Generalized Autoregressive Conditional Heteroskedasticity (First-Order Beta-Skew-T-GARCH) model. The ZAR/USD exchange rate is an important indicator for global trade, investment, and economic stability. However, traditional volatility models often struggle to fully capture its complex behaviour. This research aims to fill this gap by using FGARCH and First-Order Beta-Skew-T-EGARCH models to gain a better understanding of volatility characteristics, including long-memory effects, asymmetry, and skewness. The outcomes of this research will contribute to the refinement of models for understanding and predicting volatility in the foreign exchange markets, providing valuable implications for financial decision-makers and policy-makers.

112

Changes on students preferred learning style over a period of three years (2021 – 2023)

Mr Gezani Richman Miyambu[1], Dr Tshepo Ramarumo[1]
[1]Sefako Makgatho Health Sciences University

The purpose of the study was to check if there are changes over the years in the methods that students prefer for learning. A census was used on all the students registered for two courses who were studying research design and statistics from 2021 to 2023 at one institution of higher learning. The size of the population of interest during the three-year period of data collection was 324 and a total of 284 students responded to the survey. This study has achieved a response rate of 88%. An online survey was used for data collection for the period mentioned above. During the 3-year period most students preferred blended learning with 41.9%, 40.4% and 50% for 2021, 2022, and 2023 respectively. Online was preferred the most in 2022 with 45% but dropped in 2023. Preference of contact learning only dropped from 19.4% in 2021 to 13.7% in 2023. These are the results of the combined students during the study period. The preference of contact learning which has been a traditional learning model in the university is low. Many students have adapted to the changes which are happening in the world. With many universities adopting blended learning, this shows that the university students are ready for the change.

119

Survival analysis of patients with hypernephroma

Ms Mamelang Molaba[1], Mr Gezani Richman Miyambu[1]
[1]Sefako Makgatho Health Sciences University, Pretoria, South Africa

Hypernephroma (also referred to as Renal Cell Carcinoma) is the type of kidney cancer that starts in the lining of small tubes in the kidney. This study focuses on predicting the probability of response, survival, or mean lifetime by comparing survival distribution of hypernephroma patients. The data used in this study is from a retrospective study for thirty-three (33) patients with hypernephroma. The study specifically focused on the length of time from the day treatment started until the time at which either, some well-defined events occur, such as death or relapse to a certain condition. Statistical package, IBM SPSS Statistics 23 was used; to obtain Kaplan-Meier survival analysis. The results for this method have proven that the response to treatment of a hypernephroma patient has a major effect on their survival time. The overall comparisons table provides overall tests of the equality of survival times where one of the tests, Breslow (Generalized Wilcoxon) has a significance value of 0.034 less than 0.05, indicating statistically significant difference between the responses in the survival time. Thus, this study has developed life expectancy model or distribution fitting for patients with hypernephroma. Considering the survival time of patients with hypernephroma; it is recommended that health faculties must encourage people to donate their organs, especially kidneys in this case, so that patients get help and as a result they would have increased hypernephroma patient's survival time.

56

A shared frailty model for left-truncated and right-censored under-five child mortality data in South Africa

Dr Tshilidzi Benedicta Mulaudzi[1]
[1]University of Venda, Thohoyandou, South Africa

Many African nations continue to grapple with persistently high under-five child mortality rates, particularly those situated in the Sub-Saharan region, including South Africa. A multitude of socio-economic factors are identified as key contributors to the elevated under-five child mortality in numerous African nations. This research endeavours to investigate various factors believed to be associated with child mortality by employing advanced statistical models. This study utilizes child-level survival data from South Africa, characterized by left truncation and right censoring, to fit a Cox proportional hazards model under the assumption of working independence. Additionally, a shared frailty model is applied, clustering children based on their mothers. Comparative analysis is performed between the results obtained from the shared frailty model and the Cox proportional hazards model under the assumption of working independence. Within the scope of this analysis, several factors stand out as significant contributors to under-five child mortality in the study area, including gender, birth province, birth year, birth order, and twin status. Notably, the shared frailty model demonstrates superior performance in modelling the dataset, as evidenced by a lower likelihood cross-validation score compared to the Cox proportional hazards model assuming independence. This improvement can be attributed to the shared frailty model's ability to account for heterogeneity among mothers and the inherent association between siblings born to the same mother, ultimately enhancing the quality of the study's conclusions.

92

Recent advances in spatial statistics methods for rail networks

Ms Nomly Ngubeni[1], Prof Inger Fabris-Rotelli[2], Dr Najmeh Nakhaei Rad[2]
[1]University of Pretoria / Transnet, Pretoria, South Africa, [2]University of Pretoria, Pretoria, South Africa

This research presents a literature review of spatial analysis methodology applied to rail systems. The rail system plays a vital role in modern transportation, offering a sustainable and efficient alternative to road travel for goods, freight and passengers. This review examines the current state of spatial methods applied to rail networks, focusing on both methodological review and practical applications. Spatial statistics methods play a crucial role in the analysis of rail networks, providing insights into network performance, connectivity, and optimisation. Various spatial statistics techniques for network analysis are discussed, highlighting their relevance and effectiveness in addressing key challenges in rail network analysis. Additionally, an application of these methods on rail network planning and optimisation is presented. Finally, gaps in current research are identified and we suggest future directions for the advancement of spatial statistics methods in the field of rail network analysis.

On classes of consistent tests for the Type I Pareto distribution based on a characterisation involving order statistics

Dr Thobeka Nombebe[1], Prof James Allison[1], Prof Leonard Santana[1], Prof Joseph Ngatchou--Wandji[2]
[1]North-West University, Potchefstroom, South Africa, [2]Universite' de Rennes (EHESP) & Institut Elie Cartan de Lorraine, Nancy, France

We propose new classes of goodness-of-fit tests for the Pareto Type I distribution. These tests are based on a characterization of the Pareto distribution involving order statistics. We derive the limiting null distribution of the tests and also show that the tests are consistent against fixed alternatives. The finite-sample performance of the newly proposed tests is evaluated and compared to some of the existing tests, where it is found that the new tests are competitive in terms of powers. The paper concludes with an application to a real-world data set, namely the earnings of the 26 highest paid participants in the inaugural season of LIV golf.

175

Multinomial regression models: an applied approach to model consumer utility and preference

Mr Macdonald Phasha[1], Dr Judy Kleyn[1]
[1]University of Pretoria, Pretoria, South Africa

Individuals make choices every day, some of them might be habitual and repetitive (such as choosing which newspaper to buy, whether to take the bus, car or train to work) or once-off (such as buying a house). In many of these choice situations, the individual usually has more than one option/alternative. In making these choices, do individuals seek to fulfil either an inherent desire (satisfaction) for personal gain or altruistic interest? In the pursuit of finding the reason, we discover an underlying objective. Regardless of what the objective is, there is an underlying reason or factors that influence why an individual made a specific choice accompanied by a set of constraints. A classic example is the quantity of food one can buy at a market; an example of a constraint is income and/or budget.

In this mini-dissertation, we aim to study the choice behaviour of individuals with the help of real-life and simulated data and model the relationship between the choices that consumers make and the underlying factors that influence these choices.

A comparison of the robust zero-inflated and hurdle models with an application to maternal mortality

Mr Phelo Pitsha[1], Mr R Chiruka[1], Dr C Marange[1], Dr M Mutambayi[1]
[1]University of Fort Hare, Alice, South Africa

Background: Zero-inflated data is a significant issue in public health studies, and researchers have developed zero-inflated and hurdle models to handle zero inflation. Robust Zero-Inflated (RZI) and Robust Hurdle (RH) models are extensions of these models, offering robustness to outliers and deviations from model assumptions.

Methods: The study applied RZI and RH models to analyse Nairobi's zero-inflated count data of maternal mortality. These models were used to account for the complexity and zero inflation of the data. The models were compared in terms of AIC and BIC, and the Voung test was used.

Results: The Poisson model had the least fit, while the Negative Binomial model was better. The RZIP model is the most suitable, with the lowest AIC of 366.9 and BIC of 400.9. The RZINB and Hurdle models also perform well but are slightly less optimal. The Vuong test confirms the RZIP model's superiority and captures additional information beyond traditional models. The study found that breach delivery and teenage pregnancies significantly increase maternal mortality rates. Early teenage pregnancy increases mortality by 6% (exp(β) = 1.06, 95% CI: 1.01-1.11, p = 0.023), late teenage pregnancy by 13% (exp(β) = 1.13, 95% CI: 1.03-1.23, p = 0.010), and Assisted Deliveries have no significant impact. The findings suggest that delivery and pregnancy characteristics play a significant role in maternal mortality, particularly in teenage pregnancies.

Conclusion: The study reveals that Robust Zero-Inflated Poisson (RZIP) models accurately represent maternal mortality data in Nairobi, efficiently handling zero inflation and outliers. This model outperforms traditional Poisson and Negative Binomial models in identifying significant factors like breach deliveries and early teenage pregnancy, enhancing the capacity to create tailored treatments and improve maternal health outcomes.

117

The prevalence and spatial dynamics of housebreaking and home robbery hotspots in South Africa

Mr Gomolemo Rakale[1], Ms Moleseng CM Ramaube, Mr Sonnyboy K Manthata, Prof Solly M Seeletse
[1]Sefako Makgatho Health Sciences University, Ga-Rankuwa, South Africa

This study investigates the prevalence and underlying dynamics of housebreaking and home robbery hotspots in South Africa during the 2017/2018 period. Drawing on the data from the victims of crime surveys compiled by Statistics South Africa, the study aims to identify and analyse crime hotspots by province. By utilising advanced spatial statistical methodologies, the study will map the geographic distribution of these hotspot areas across all provinces, uncovering patterns of occurrence that may be linked to broader socio-economic and environmental factors. The study will identify areas of priority for intervention and prevention of housebreaking and home robbery by exposing locations where most of these crimes occur, and the times and days in which they mostly occur.

Comparison of the discrete-time survival model and machine learning models

Mr Audrey Tshepho Ramachela[1]
[1]University of Venda, Moletjie, South Africa

Prediction models for survival analysis are commonly used in biomedical sciences to understand the onset of certain diseases. Traditional statistical models have been employed in the past; however, their limitations and inability to handle big data sets have made a way for the introduction of machine learning methods which gained recognition due to their ability to learn complex patterns. However, existing literature indicates that the predictive accuracy of machine learning and statistical models for survival analysis varies significantly across different data sets. This variability underscores the need for further research utilizing data sets with diverse characteristics. Such research is essential to develop generalizable insights into the conditions under which each method performs best. In this paper, we propose to compare the predictive performance of traditional statistical methods and machine learning algorithms in discrete survival analysis. The machine learning methods include survival trees, bagging-survival trees, random survival forests, and neural networks. The study uses calibration (measured by the Brier score) to assess model fit, and discrimination (measured by the Concordance index and area under the curve) to evaluate predictive accuracy. These methods are applied to data sets on Copenhagen Stroke, breast cancer and age at first alcohol intake. The results computed from the Copenhagen Stroke study data set suggest that a tree fitted through the Ranger package in R has the best overall performance, followed by discrete-random survival forest. The results also show that the machine learning algorithms have better prediction performance as compared to the traditional statistical models.

115

Work integrated learning challenges in a specific academic department of a Gauteng-based University

Dr Tshepo Ramarumo[1], Mr Gezani Richman Miyambu[1]
[1]Sefako Makgatho Health Sciences University, Ga-Rankuwa, South Africa

Universities are currently introducing work integrated learning (WIL) programme. The WIL programme is well functioning in the universities of technologies, which are the former Technikons, as they were focusing on work-related programmes. During Technikons days, the study programmes were 50:50 theory and practical. The practical part was what now takes the space of WIL. The universities did not have that option, and their notational hours required 100% theory. Introducing a new programme in the module is challenging. The purpose of this study was to investigate challenges encountered from the Statistical Sciences departmental members as well as the organizer when the programme was introduced. This study used a qualitative approach to enable an in-depth exploration to understand the challenges in the Statistical Sciences department. The discussion demonstrated the importance of training (workshops) when introducing a programme.

11

Predicting the closing price of cryptocurrency Ethereum

Dr Thakhani Ravele[1], Mr Vhukhudo Ronny Rambevha[1], Prof Caston Sigauke[1]
[1]University of Venda, Thohoyandou, South Africa

Given that cryptocurrencies are now involved in nearly every financial transaction due to their widespread acceptance as an alternative method of payment and currency exchange, researchers and economists have increased opportunities to analyse cryptocurrency prices. Over time, predicting the daily closing price of Ethereum has been challenging for investors, traders, and investment banks because of its significant price volatility. The daily closing price of cryptocurrency is crucial for trading or investing in Ethereum. This report aims to conduct a comparative analysis of the predictive performance of deep machine learning algorithms within a stacking ensemble modelling framework, utilizing daily historical price data of Ethereum from Coindesk, tweets from Twitter spanning from August 1, 2022, to August 8, 2022, and five additional covariates (closing price lag1, closing price lag2, noltrend, daytype, and month) derived from Ethereum's closing price. Seven models are employed to forecast the daily closing price of Ethereum: recurrent neural network (RNN), ensemble stacked recurrent neural network, gradient boosting machine, generalized linear model, distributed random forest, deep neural networks, and a stacked ensemble of gradient boosting machine. The primary evaluation metric is the mean absolute error (MAE). Based on MAE, the RNN forecasts outperform the other models in this study, achieving an MAE of 0.0309.

149

Modelling the probability of default using logistic regression and threshold-logistic regression

Miss Monalisa Williams[1], Miss Lesego Sepato[1]
[1]Nelson Mandela University, Port Elizabeth, South Africa

The study investigates the driving factors contributing to loan defaults within a consumer finance company that specialises in lending various types of loans to urban customers. The company wants to understand the key variables (factors) and identify patterns that differentiate “bad loans” from “good loans” for improved decision-making process for loan approval. The study compared logistic regression and threshold regression models to assess default risk based on applicant profiles. Factors such as high interest rates, low credit scores, high debt-to-income ratios (DTI), and increased inquiries in the last six months are shown to significantly contribute to higher default risk. Additionally, the analysis reveals that loan term, loan purpose, and loan size also influence default probabilities. Smaller loan amounts show a higher risk of default, although this risk levels off beyond a certain loan size.

Results indicate that while logistic regression achieves high accuracy (AUC = 70.87%), it struggles with identifying defaulters in imbalanced datasets. In contrast, the threshold regression model (AUC = 70.94%) provides a better balance between sensitivity and specificity, making it more effective at predicting defaults, particularly when risk factors change at specific thresholds.

Key findings suggest that interest rates, credit score, DTI, revolving utilisation, and loan term are strong predictors of default. Moreover, small business loans and education loans exhibit higher default rates, while debt consolidation is the most common loan purpose. The study concludes that threshold regression is the more robust model for predicting defaults in varying loan conditions, offering critical insights for better risk management and loan approval strategies.

A new alpha power Weibull model for analysing time-to-event data: application to diabetes mellitus data

Assistant Prof Getachew Tekle[1], Dr Gao Shengjie, Dr Alisa Craig
[1]Wachemo University, Hossana, Ethiopia

The development and introduction of new families of probability distributions have been an additional interest for the expertise in applied statistics. The adaptation, extension, modification and application of additional parameters to the existing probability distribution models are some of the techniques commonly used. Statistical methodologies have wider applications in exercise science, sports medicine, sports marketing, sports science, and other related sciences. These methods can be used to predict the winning probability of a team or individual in a match, the number of minutes that an individual player will spend on the ground, the number of goals to be scored by an individual player, the number of red or yellow cards that will be issued to an individual player or a team, etc. Keeping in view the importance and applicability of the statistical methodologies in sport sciences, healthcare, and other related sectors, this paper introduces a novel family of statistical models called new alpha power family of distributions in the health sector. It is shown that numerous properties of the suggested method are similar to those of the new Weibull-X and exponential type distributions. Based on the novel method, a special model, namely, a new alpha power Weibull distribution is studied. The new model is very flexible and a good alternative to the exponential type because the shapes of its probability density function can either be right-skewed, decreasing, left-skewed, or increasing. Furthermore, this new distribution is also able to model real phenomena with bathtub-shaped failure rates. Finally, the applicability of the proposed distribution is shown by analysing the time-to-event datasets selected from public health and it is compared to three popular standard probability distributions (based on four model adequacy measures, and it is selected as the best model). The newly proposed novel model is the best model to fit the data.

145

Multivariate techniques application to reveal mutual trends among data sources: a consumer research case study

Mrs Marieta van der Rijst[1], Ms Marbi Schwartz[2], Dr Jeannine Marais[2]
[1]Agricultural Research Council, Stellenbosch, South Africa, [2]Stellenbosch University, Stellenbosch, South Africa

In consumer research there are different sources of information that need to be linked to each other, for example product descriptors, consumer preference of the products, and consumer perceptions. Identification of relationships between these sources of information is the key to successful consumer research. However, it often is a challenge to sensibly combine different types of consumer attributes to obtain the best possible interpretation. This paper explores a study that was conducted using a combination of consumer perceptions and physical measurements to evaluate internal meat colour change with internal end-point temperature change during cooking. Consumers were asked to categorise 15 randomly presented photographs into one of five perceived degree of doneness (DOD) categories, to indicate the photograph most preferred based on internal meat colour, as well as their general DOD preferences for beef steak. Physical measurements were conducted on individual steaks. Principal Component Analysis (PCA) was conducted to elucidate the association between the actual end-point internal temperatures of the steaks in the photographs and the perceived DOD. Furthermore, DOD preference was projected onto the PCA bi-plot as supplementary variables to investigate the agreement between indicated preference category and classification of preferred photograph, without influencing the association between actual end-point internal temperatures and perceived DOD. To verify observed groupings of photographs PCA factor scores were subjected to agglomerative hierarchical clustering. Resourceful utilization of standard multivariate techniques may support improved interpretation capability.

131

Modelling and forecasting headline inflation in Ethiopia by supervised machine learning approach

Mr Teklu Nega Yimenu[1], Assistant Prof Dereje Danbe[2], Dr Zeyitu Asfaw[3]
[1]Oda Bultum University, Chiro, Ethiopia, [2]Hawassa University, Hawassa, Ethiopia, [3]Addis Ababa University, Addis Ababa, Ethiopia

Inflation is an important indicator of a nation's welfare and has become one of the major economic challenges globally, especially in Ethiopia. Several studies forecasted inflation in Ethiopia using traditional models, as accurate forecasts contributed to a more stable economic environment, even though forecasting with traditional models was challenging. Thus, this study aimed to model and forecast headline inflation in Ethiopia using a machine learning approach. The study was based on secondary data obtained from various inflation-related organizations. The data were transformed, standardized and split into training and testing sets to enhance the forecast accuracy of both the machine learning. The selected models were evaluated based on performance evaluation criteria, including RMSE, MAE, and MAPE tests. The finding revealed that food inflation, non-food inflation, export and import prices of goods and services, political stability index, exchange rate, numbers of vehicles, rainfall, world oil price, gross domestic fixed investment, unemployment rate, T-bill sales, and agricultural production price were predictors which significantly determine the headline inflation in Ethiopia. Among various forecasting methods, a specific ANN architecture called Non-linear Neural Autoregressive out-performed Ridge regression, LASSO, Elastic Net, Random Forest, and even the benchmark model in terms of accuracy for both in-sample and out-of-sample inflation forecasts in Ethiopia with lowest RMSE, MAE and MAPE. Finally, this study recommended that policymakers, financial analysts, investor and stakeholders should give attention to the identified drivers of headline inflation and consider using advanced machine learning models.