NITheCS
National Institute for Theoretical and Computational Sciences

SARChI Research Chair in Biostatistics

ZA Hub
BMW digital & global solutions

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

INTERNATIONAL SYMPOSIUM ON MODERN STATISTICS AND BIOSTATISTICS (SYMSTAT2024)

26 – 28 JUNE 2024

FUTURE AFRICA

UNIVERSITY OF PRETORIA

PROGRAM

Wednesday 26 June

08:00-08:30	Arrival & Registration	VENUE	SESSION CHAIR
08:30-08:50	OPENING Official Opening Prof Barend Erasmus: Dean of the Faculty of Natural and Agricultural Sciences, University of Pretoria Welcoming Remarks Prof Samuel Manda: Head of the Department of Statistics Welcoming Remarks Prof Din Chen: SARChI Research Chair in Biostatistics	Auditorium	Dr Nakhaeirad
08:50-09:40	Plenary: Dr Yende-Zuma	Auditorium	Dr Nakhaeirad
09:40-10:05	*Mr AT Kelbrick*	Auditorium	Prof Millard
	Mr A Ngwira	Conference 1	Dr Thiede
	Dr MJC Malela	Conference 2	Dr Malela
10:05-10:30	*Ms BLM McCall*	Auditorium	Prof Millard
	Dr RN Thiede	Conference 1	Dr Thiede
	Prof A Bekker	Conference 2	Dr Malela
10:30-11:00	Tea & Snacks		
11:00-13:00	Workshop 1: Prof Samadi	Auditorium	Mrs Stander
13:00-14:00	Lunch		
14:00-14:50	Plenary: Prof Asgharian	Auditorium	Dr Nakhaeirad
14:50-15:15	Tea		
15:15-15:40	*Ms L Potgieter*	Auditorium	Prof Fabris-Rotelli
	Mr MW Waja	Conference 1	Prof Din Chen
	Dr TM Kaombe	Conference 2	Prof Manda
15:40-16:05	*Ms J Nel*	Auditorium	Prof Fabris-Rotelli
	Dr HM Fenta	Conference 1	Prof Din Chen
	Dr BA Ejigu	Conference 2	Prof Manda
16:05-16:30	*Mrs M de Klerk*	Auditorium	Prof Fabris-Rotelli
	Dr DB Belay	Conference 1	Prof Din Chen
	Prof SOM Manda	Conference 2	Prof Manda
16:30-16:55	Dr MA Lakeh	Conference 1	Prof Din Chen

Thursday 27 June

08:00-08:30	Arrival & Registration	VENUE	SESSION CHAIR
08:30-09:15	Plenary: Dr Reddy	Auditorium	Prof Fabris-Rotelli
09:15-09:40	*Mr AR Kleynhans*	Auditorium	Prof Millard
	Prof IN Fabris-Rotelli	Conference 1	Prof Fabris-Rotelli
	Mr LO Baloi	Conference 2	Prof Din Chen
09:40-10:05	Mr SB Skhosana	Auditorium	Prof Millard
	Mrs R Stander	Conference 1	Prof Fabris-Rotelli
	Ms AT Mberi	Conference 2	Prof Din Chen
10:05-10:30	*Mr AF Otto*	Auditorium	Prof Millard
	Mr K Mahloromela	Conference 1	Prof Fabris-Rotelli
	Ms NB Mashinini	Conference 2	Prof Din Chen
10:30-11:00	Tea & Snacks		
11:00-13:00	Workshop 2: Dr van Niekerk	Auditorium	Mrs Stander
13:00-14:00	Lunch		
14:00-14:50	Plenary: Prof Samadi	Auditorium	Dr Nakhaeirad
14:50-15:15	Tea		
15:15-15:40	*Mrs L Venter*	Auditorium	Prof Din Chen
	Mr RW Greyling	Conference 1	Prof Kanfer
	Mr PS Mdletshe	Conference 2	Dr Thiede
15:40-16:05	*Mr GC Singini*	Auditorium	Prof Din Chen
	Mr RH Coetzee	Conference 1	Prof Kanfer
	Mr A Antonio	Conference 2	Dr Thiede
16:05-16:30	*Mr CM Jardim*	Auditorium	Prof Din Chen
	Mr AK Krishnannair	Conference 1	Prof Kanfer
	Ms M Steyn	Conference 2	Dr Thiede
16:30-16:55	Dr N Nakhaeirad	Auditorium	Prof Din Chen

Friday 28 June

08:00-08:30	Arrival & Registration	VENUE	SESSION CHAIR
08:30-09:40	STUDENT INFORMATION SESSION SARChI Bursary Prof Din Chen: SARChI Research Chair in Biostatistics University of Pretoria Masamu Program Prof Overtoun Jenda: Auburn University, Alabama, United States of America	Auditorium	Prof Fabris-Rotelli
09:40-10:05	Prof DC Janse van Rensburg	Auditorium	Dr van Staden
	Dr AA Ashiagbor	Conference 2	Prof Kanfer
10:05-10:30	Dr PJ van Staden	Auditorium	Dr van Staden
	Dr L Chaka	Conference 2	Prof Kanfer
10:30-11:00	Tea & Snacks		
11:00-13:00	Workshop 3: Prof Sharp		
13:00-14:00	Lunch		
14:00-14:50	Plenary: Prof Din Chen	Auditorium	Prof Fabris-Rotelli
14:50-15:15	Closing & Prize Giving	Auditorium	Prof Fabris-Rotelli

STREAM LABELS
Plenary
Workshop
Machine Learning
Spatial Statistics
Statistical Methodology
Biostatistics
Sports Statistics

WIFI DETAILS

Steps to login:

1. **Click on/ select Tuks Guest wifi**
2. **Where it says "already have an account" click on sign in.**
3. **Use details attached to login and accept the T&C's**
4. **It should direct you to the UP website and then log you in.**
5. **Guest username: biostatistics@up.ac.za, Guest Password: 8695**

ABSTRACTS

PLENARY SPEAKERS

Professor Masoud Asgharian: Professor in Statistics, Department of Mathematics and Statistics, McGill University, Canada

Prevalent Cohort Studies: Length-Biased Sampling with Right Censoring

Logistic or other constraints often preclude the possibility of conducting incident cohort studies. A feasible alternative in such cases is to conduct a cross-sectional prevalent cohort study for which we recruit prevalent cases, that is, subjects who have already experienced the initiating event, say the onset of a disease. When the interest lies in estimating the lifespan between the initiating event and a terminating event, say death for instance, such subjects may be followed prospectively until the terminating event or loss to follow-up, whichever happens first. It is well known that prevalent cases have, on average, longer lifespans. As such, they do not form a random sample from the target population; they comprise a biased sample. If the initiating events are generated from a stationary Poisson process, the so-called stationarity assumption, this bias is called length bias. My work revolves around developing statistical methodologies for analyzing such data. Our study is mainly motivated by challenges and questions raised in analyzing survival data collected on patients with dementia as part of a nationwide study in Canada, called the Canadian Study of Health and Aging (CSHA). I'll use these and other real data for my work to discuss and motivate our methodologies and their applications.

Professor Yaser Samadi: Associate Professor of Statistics, School of Mathematical and Statistical Sciences, Southern Illinois University Carbondale, USA

Multivariate Time Series Analysis through Reduced-Rank Envelope Vector Autoregressive Models

Vector autoregressive (VAR) models have historically been favored for their adaptability and simplicity in modeling multivariate time series data. However, the VAR framework often encounters overparameterization issues, particularly in high-dimensional time series datasets, limiting the incorporation of variables and lags. Several statistical approaches have been proposed to address dimension reduction in VAR models, yet, they prove inefficient in extracting relevant information from complex datasets, as they fail to distinguish between information aligned with scientific objectives and are also inefficient in addressing rank deficiency problems. In this context, envelope methods offer a promising solution by leveraging reduced subspaces to identify and eliminate irrelevant information, thereby enhancing efficiency in parameter estimation. This presentation introduces an innovative VAR model integrating envelope concepts within the reduced-rank framework, facilitating substantial dimension reduction without compromising parameter estimation accuracy. Through comprehensive simulation studies and real-world data analysis, we demonstrate the superior performance of our model compared to existing methodologies in the literature, underscoring its efficacy in capturing essential dynamics while mitigating the limitations of traditional VAR frameworks.

Professor Ding-Geng (Din) Chen: Executive Director and Professor in Biostatistics at the College of Health Solutions, Arizona State University; Extraordinary Professor and the SARChI Research Chair in Biostatistics at the University of Pretoria; Honorary Professor at the University of KwaZulu-Natal

How to Estimate COVID-19 Vaccine Efficacy

The COVID-19 pandemic has caused significant morbidity and mortality, as well as social and economic disruption worldwide. In order to reduce these effects, a global effort to develop effective vaccines against the COVID-19 virus has produced various options with the effectiveness assessed on the rate of infection between vaccinated and unvaccinated groups, which has been used for important policy decision-making on vaccination effectiveness ever since. However, the rate of infection is an over-simplified index in assessing the vaccination effectiveness overall, which should be strengthened to address the duration of protection with time-to-infection effect. The fundamental challenge in estimating the vaccination effect over time is that the time-to-infection for unvaccinated group is unknown due to nonexistent vaccination time. This presentation is to discuss the biostatistical methodological development to fill this knowledge gap to propose a Weibull regression model. This model treats the nonexistent vaccination time for the unvaccinated group as nuisance parameters and estimate the vaccination effectiveness along with these nuisance parameters. The performance of the proposed approach and its properties is empirically investigated through a simulation study, and its applicability is illustrated using a real-data example from the Arizona State University COVID-19 serological prevalence data.

Dr Nonhlanhla Yende-Zuma: Specialist Statistician, South African Medical Research Council (SAMRC)

Causal inference methodology in the context of future HIV prevention clinical trials

Randomised controlled trials (RCTs) remain the gold standard for evaluating the efficacy of new or emerging interventions. In the context of HIV prevention, we discuss that the use of a placebo as a comparator is becoming unethical and requires justification as highly effective long-acting injectable pre-exposure prophylaxis (PrEP) agents become available. On the other hand, active-controlled trials will require enormous sample sizes and probably prohibitive costs.

We provide alternative approaches that could be utilised in designing future HIV prevention trials, such as non-inferiority design, use of registrational cohorts, non-randomised comparator groups such as historical placebo controls and immune biomarkers as a mediator of prevention efficacy. However, these approaches could produce biased efficacy estimates due to potential confounding.

Most importantly, we discuss causal inference models such as propensity scores, instrumental variables, marginal structural models, etc, as they have the potential to provide robust estimates in this context.

Dr Tarylee Reddy: Biostatistics Unit Director, South African Medical Research Council (SAMRC)

Time to threshold estimation from longitudinal biomarker data: continuous, censored, discrete data and beyond

In longitudinal studies of biomarkers, an outcome of interest is the time at which a biomarker reaches a particular threshold. Due to the inherent variability of several studies have applied persistence criteria, designating the outcome as the time to the occurrence of two consecutive measurements less (or greater) than the threshold. In this presentation, we discuss a method to estimate the time to attainment of two consecutive measurements less than a meaningful threshold, which takes into account the patient-specific trajectory and measurement error. An expression for the expected time to threshold has been presented, which is a function of the fixed effects, random effects and residual variance. While the initial approach was motivated by continuous biomarkers, we present extensions of the methodology to accommodate censored observations as well as ordinal outcomes. We present a range of specific applications to HIV, cardiology, SARS-CoV2 and schizophrenia demonstrating the relevance of the methodology. Through these applications we demonstrate that the method proposed is computationally efficient, robust, and offers more flexibility than existing frameworks.

CONTRIBUTED TALKS

Mr A Antonio (University of Pretoria)

Co-author(s): Prof IN Fabris-Rotelli (University of Pretoria), Dr RN Thiede (University of Pretoria), Mrs R Stander (University of Pretoria)

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

This study quantifies the overlap between existing police precinct boundaries and theoretically optimal boundaries derived from Voronoi diagrams based on Euclidean and network distances. Spatial similarity measures are used to analyse the relationship between boundary overlap and police station accessibility, hypothesizing that reduced overlap corresponds to decreased accessibility. Additionally, the potential impact of boundary placement on crime rates is explored, suggesting a correlation between less accessible police stations and increased crime levels. By quantifying these relationships, this research aims to evaluate the effectiveness of current precinct boundaries and their potential influence on crime.

Dr AA Ashiagbor (Department of Business Management, University of Pretoria)

Co-author(s): Prof S. Das (Department of Business Management, University of Pretoria)

Transmuted Claim-C Family of Distributions with Applications to Cancer Disease Data

The lifetime distribution plays an important role in many real-life fields, such as biostatistics, reliability, and survival analysis. We are motivated to contribute to finding new lifetime distributions and introduce a new family of continuous distributions called the Transmuted Claim-C (TCC) family, which is a flexible and robust extension of the Claim-C distribution. The TCC family is defined using a transmutation map, which allows for the modeling of complex and asymmetric data. We develop two special models based on the Weibull and Burr Type III distributions. Various properties and characteristics of the TCC family, including its probability density function, cumulative distribution function, quantile function, and moment generating function, are explored. We also present estimation methods and applications of the TCC family to real-world data, specifically in the context of cancer disease modeling. Our results show that the TCC family provides a better fit to cancer data compared to existing distributions based on the Akaike information criterion and some goodness-of-fit statistics, highlighting its potential for use in medical research and clinical applications. The proposed TCC family offers a valuable tool for modeling and analyzing complex data in various fields, including medicine, engineering, and the social sciences.

Mr LO Baloi (Department of Statistics, University of Pretoria)

Co-author(s): Dr N Nakhaeirad (Department of Statistics, University of Pretoria), Prof D Chen (Department of Statistics, University of Pretoria)

Estimation of incubation period and generation time of COVID-19 based on observed length-biased and interval censored epidemic cohort.

The COVID-19 pandemic has underscored the importance of accurately estimating the incubation period and generation time of infectious diseases, both critical parameters for effective epidemiological modeling and public health decision-making. The incubation period, defined as the interval between infection and symptom onset, is crucial for determining optimal quarantine durations. Generation time, the period between the infection of a primary case and secondary cases, influences estimates of the basic reproduction number, a key measure of disease transmissibility. To estimate the incubation distribution, the most common approach is contact-tracing–based estimation which is highly impacted by people's assessment about the possible date of exposure and this might cause significant errors. The alternative Interval censoring–based methods are able to handle big data but may face biased sampling. Observed serial intervals are usually applied to estimate the distribution of generation time, but when the disease is infectious during incubation, it may provide a biased approximation. In this research, a dataset consisting of departure time from Wuhan and onset of COVID-19 symptoms for 1211 passengers is analyzed. We define the incubation period as the interarrival time, and the duration between departure and onset of symptoms as the mixture of forward time and interarrival time with censored intervals. The distribution of incubation is estimated through theory of renewal process and interval censoring with a mixture distribution. In addition, a consistent estimator for the distribution of generation time is obtained using incubation period and serial interval for incubation-infectious diseases.

Prof A Bekker (University of Pretoria)

From Data to Wisdom: A playground of innovations in statistical learning

In this research we provide superior solutions where complex systems and data sets are present in multidisciplinary scientific spectra using statistical and probabilistic learning models. Developments in network learning and graphical models, machine learning and extended distribution theory for manifolds form the basis. The talk will encompass modelling innovations with the impact visible on datasets from a variety of areas, including biomechanical engineering, sustainable energy, and financial risk management. It is evident that sustainability science is closely connected with sophisticated data analysis. This research will explore new pathways to unlock the hidden truths that lie in the data.

Dr DB Belay (University of Pretoria)

Co-author(s): Prof Ding-Geng Chen (UP), Dr Najmeh Nakhaeirad (UP)

Bayesian Multivariate joint modeling of longitudinal and time-to-event data using HIV/AIDS patients

HIV/AIDS is still a global health burden and causes a significant number of morbidity and mortality. The application of joint modelling is a statistical approach that simultaneously analyses multiple longitudinal outcomes and time-to-event data which recently became crucial in HIV/AIDS research. This study aims to quantify the risk of time-dependent biomarkers and the time to death of HIV/AIDS patients using the Bayesian joint model estimation technique. About 571 patients' registration cards were accessed and their information is recorded from Felege Hiwot Hospital, Ethiopia. Out of 571 patients, about 55.2% and 44.8% are female and male participants, respectively. This study shows that the longitudinal measures of CD4 cell counts and weight of patients are significantly associated with the time to death of patients at different functional forms and association structures. The hazard of death of the patients decreased by 13% as a unit increase in squared root transformed CD4 cell count. Similarly, the unit increase in a logarithm of the weight of the patients can decrease the hazard of the patient by 98.3% which shows an instantaneous effect on time to death of patients. Using longitudinal and time-to-event data jointly gives a more robust estimation and will help the true patient-specific intervention. In this study, we analyzed the two longitudinal biomarkers on time to death of HIV/AIDS patients which led to the new hypothesis that the patient’s weight should be considered as an important biomarker as CD4 cell counts to determine the progress of the patient’s health status.

Dr L Chaka (University of Pretoria)

Co-author(s): Prof P.M. Njuho (University of Zululand, South Africa), Prof H-.P. Piepho (University of Hohenheim, Germany)

A new approach to linear mixed models when each factor consists of both fixed and random levels.

The era of big data and its complexities has impacted the field of data analytics, leading to the emergence of some crucial innovations in technology and data analysis strategies in the past decade. Recent technological advancement in fields such as agriculture and other industrial processes has led to the development of a statistical modelling concept 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 conceptualized 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 structure of the variance-covariance matrix of error terms. 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 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 researcher's choice and the objectives of the experiment. However, combining these partitioned analyses of variance for a broader perspective poses some challenges on re-arrangement of the data and coding prior to analyses where statistical software has no direct provision that handles such complexity. In addition, SAS and CRAN R environments are proposed and used to obtain the combined analyses.

Mr RH Coetzee (University of Pretoria)

Co-author(s): Prof SM Millard (University of Pretoria), Prof FHJ Kanfer (University of Pretoria)

Non-parametric modal regression

Modal regression is a data analysis tool that may be used alongside or alternatively to traditional approaches that regress towards the mean. It offers robustness to outliers and is particularly effective for skewed distributions. This presentation focuses on kernel-based non-parametric modal regression, which aims to model conditional modes without imposing strong assumptions on the data. The presentation begins by highlighting the significance of the mode as a useful summary statistic, representing the most likely value in a distribution. A key aspect of the methodology involves the utilisation of the General Modal Expectation-Maximization (GMEM) algorithm, which leverages kernel density estimation to find conditional modes efficiently. A link is drawn between kernel density estimation and mixture distributions, facilitating the accurate modelling of conditional modes. A simulation study is conducted to evaluate the performance of the proposed modal regression model against existing methods. We also present an application of the modal regression model to real-world data, showcasing its practical utility in data analysis tasks.

Mrs M de Klerk (University of Pretoria)

Co-author(s): Prof I Fabris-Rotelli (University of Pretoria)

Hospital accessibility catchment areas as a fuzzy lattice data structure

The accessibility to basic facilities and services plays a pivotal role in every society and city planning. Spatial accessibility can vary between cities and countries and is mainly defined by the ease at which facilities can be accessed by communities. Facilities can provide essential services and/or products such as pharmacies, clinics, schools, universities, etc. Spatial accessibility is dependent on the spatial impedance between a facility and the target population and can be illustrated with catchment areas. We propose a fuzzy lattice catchment area method which uses a semi-supervised learning algorithm to create overlapping catchment areas. This methodology is applied to determine the accessibility to hospitals in South Africa and provides an illustration on the difference for regions with high accessibility compared to low accessibility. The application can easily be adapted in a variety of fields based on industry type, drive-time thresholds, supply capacity and the target population.

Dr BA Ejigu (Department of Statistics, Addis Ababa University)

Co-author(s): Prof Samuel Manda (Department of Statistics, University of Pretoria, South Africa)

Covariate-based Nonstationary Geostatistical Modeling Framework: Spatial prediction based on first and third law of geography

The standard model-based geostatistics assume stationarity and isotropy but fail to capture the variation in statistical characteristics of the study variable across a study area. Furthermore, spatial prediction based on this modeling approach is done based on the concept of the First Law of Geography and/or statistical principles. However, depending on the nature of the process under study, two locations that are close geographically but separated by other factors may not be considered as near neighbors (i.e., only the First Law of Geography may not fully capture spatial dependence).

In this talk, I will discuss recent developments to model the non-stationary pattern of the data by incorporating covariates in the covariance structure of the model to construct non-stationary processes. Additionally, a new approach to spatial prediction using both the concept of First and Third Laws of Geography will be discussed. This approach allows for high-quality predictions without requiring large sample sizes and is particularly valuable in disease control, where accurate predictions are crucial for effective interventions. The practical utility of considering covariate(s) in the covariance structure of the model to handle non-stationarity will be demonstrated by analyzing malaria in Mozambique, anemia in Ethiopia, and HIV in South Africa.

Prof IN Fabris-Rotelli (University of Pretoria)

Co-author(s): Ms Tshepiso Rangongo (University of Pretoria), Dr Renate Thiede (University of Pretoria)

Assessing classification performance for sampled remote sensing data

Big data poses challenges for storage, management, processing, analysis and visualisation. One technique of handling big data is the use of a representative sample of the data. This paper proposes a sampling algorithm which makes use of multivariate stratification with the aim of obtaining a sample that best represents the population while minimising the number of images in the sample. The proposed sampling algorithm performs effectively on a big spatial image dataset of crop types. The results are assessed by measuring the number of images sampled and as well as matching the proportionality of the population crop percentages. The samples obtained from the proposed algorithm are then used for land cover classification. An ensemble method called random forest is trained on the samples and accuracy is assessed. Precision, recall and F1-scores per crop type are computed as well as the overall accuracy. The random forest classifier performed best on the proposed sample with the least number of images. In addition, the classifier performed better on the proposed sample than it did on a random sample as the proposed sample due to the more informative data. This research develops an effective way of sampling big data for crop classification.

Dr HM Fenta (University of Pretoria)

Co-author(s): Prof Ding-Geng Chen (University of Pretoria), Prof Temesgen T Zewotir (university of KwaZulu-Natal), Dr Najmeh Nakhaei Rad (University of Pretoria)

Spatiotemporal Models with Confounding Effects: Application on Under-Five Mortality across sub-Saharan African Countries

Under-five mortality (U5M) in sub-Saharan African (sSA) countries is a major child health concern. The main aim of this study is to adopt the spatiotemporal dynamic model which includes the confounding effects between time, space, and their interactions with fixed covariates, with special emphasis on U5M across disconnected sSA countries. We used the nationally publicly representative Demographic and Health Survey data for the period from 2000 to 2020. Bayesian spatiotemporal hierarchical modeling with an integrated nested Laplace approximation program was used to model the spatiotemporal distribution of U5M among children across 37 districts located in four disconnected sSA regions, consisting of Ethiopia, Nigerian, Zimbabwe, and Ghana. A total of 170,356 U5 children from 37 districts were considered and 15,467 died before the age of five. The relative risk of U5M in the first DHS was 2.02 and it sharply decreased to 0.5 in the recent phase. The proportion of improved access to water, sanitation, clean fuel, urbanization, and access to health facilities in the district had a significant negative association with U5M: the higher the proportion of these covariates, the lower the prevalence of childhood mortality. This study revealed the evidence of strong spatial, temporal, and interaction effects influencing the under-five mortality risk across the districts. Improving the women's literacy index, access to improved water, and the wealth index are associated with the improvement of risk of mortality among under-five children across the districts. Special attention should be paid to districts found in Nigeria and Ethiopia.

Mr RW Greyling (University of Pretoria)

Co-author(s): Prof SM Millard (UP), Prof FHJ Kanfer (UP)

An unsupervised K-means algorithm for clustering

The traditional K-means algorithm requires the pre-initialization of the number of clusters K. This process leads to the necessity to rerun the algorithm multiple times to find the optimal value of K, which makes the process inefficient. Our research considers the work by Sinaga and Yang (2020) developing a K-means algorithm that does not require a pre-specified value for K. The algorithm is implemented on both simulated and real datasets to demonstrate the applicability of the algorithm.

Prof DC Janse van Rensburg (Section Sports Medicine, Faculty of Health Sciences, University of Pretoria)

Co-author(s): Dr Audrey Jansen van Rensburg (Section Sports Medicine, Faculty of Health Sciences, University of Pretoria)

Managing Travel- and Jetlag-Induced Physical Disturbances in Elite and Recreational Athletes

For the modern-day elite athlete, intercontinental travel is a frequent occurrence. As opportunities for international competition and training increase, travel is also the reality for many recreational level or sub-elite athletes. A high volume of travel and transcontinental crossing can result in complicated physical disturbances and symptoms. Jetlag and travel fatigue pose challenging problems for the athlete, and rapid eastward or westward travel may negatively affect the body and hamper athletic performance. The human circadian phase that gets disrupted with jetlag is elusive and thus limits evidence-based therapeutic advice. A better understanding of pre-flight, in-flight, and post-flight management possibilities is important for the sports physician, the coach, the athlete, and, indeed any traveller. This session will critically review available literature and provide practical suggestions for implementing a travel management program.

Mr CM Jardim (University of Pretoria)

Co-author(s): Prof I Fabris-Rotelli (University of Pretoria), Dr A de Waal (University of Pretoria), Dr N Nakhaei Rad (University of Pretoria)

Modelling protein dihedral angles on the torus

We present a comprehensive approach to modelling bivariate toroidal data through the application of more flexible distributions and kernel density estimation techniques. Dihedral angles, pivotal in defining protein conformation, are one of the 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 behavior of proteins. This improvment in computational methods and structural understanding can lead to advancements in protein engineering and improve drug design.

We perform an in-depth analysis on kernel density estimation techniques and mixture-based density estimation techniques utilising flexible component distributions. Our methodology includes robust parameter estimation using optimisation techniques such as differential evolution. Cross-validation and other practical approaches are used for selecting bandwidth parameters. We explore and compare these approaches on protein dihedral angle data to provide a detailed understanding of the efficacy of our approaches.

Dr TM Kaombe (Department of Mathematical Sciences, University of Malawi

Co-author(s): Prof Samuel O.M. Manda (Department of Statistics, Faculty of Natural and Agricultural Sciences, University of Pretoria, Pretoria, South Africa)

Influence Diagnostic Statistics for Multivariate Survival Data Modelling

Measures for the identification of influential observations are well-established in linear models. However, there is very little research on identifying influential groups of observations in the analysis of multivariate survival data. In this paper, we extend the martingale-based residuals and leverage commonly used in univariate survival regression to derive an influence measure for multivariate survival data models. The performance of the proposed diagnostic measure is evaluated using Monte Carlo simulation studies. Additionally, we demonstrate its usefulness with an analysis of clustered child survival data to identify influential clusters of children's survival observations.

Mr AT Kelbrick (Department of Statistics, University of Pretoria)

Co-author(s): Dr N Nakhaeirad (Department of Statistics, University of Pretoria), Dr PJ van Staden (Department of Statistics, University of Pretoria), Prof V Maharaj (Department of Chemistry, University of Pretoria)

Virtual screening of plants for anti-malarial activity using machine learning

Many drugs used today were inspired by nature. Discovering active compounds within plants that target specific diseases, paves the way for developing new medicines. This presentation outlines the use of word embedding-based virtual screening (WEBVS) to predict novel compounds from plants for anti-plasmodial activity. 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. Bio-assay has been the main method employed, which is a time consuming and often manual process. The WEBVS aims to supplement this approach by using available textual data from literature. In this presentation, the abstracts of over 400 000 papers were used to generate the word embeddings. A continuous bag-of-words (CBOW) architecture was used to encode literature on plants. A neural net with dropout regularization was then trained on a hand-crafted training set to classify plants into either active or inactive categories. Labels were based on whether or not a plant extract contained a compound with a half-maximal inhibitory content (IC 50) value of less than 1 000 against a quinine resistant strain of plasmodium falciparum. Our model achieved a 95.29% k=5 fold cross validated binary accuracy. A shortlist of the top predicted plants and compounds were assessed using bio-assay.

Mr AR Kleynhans (University of Pretoria)

Co-author(s): Prof SM Millard (University of Pretoria), Prof FHJ Kanfer (University of Pretoria)

A component self-paced learning algorithm for fitting finite mixture models

The self-paced learning (SPL) algorithm estimates the parameters of a finite mixture model (FMM) by introducing observations in a meaningful order, considering each observation’s contribution to the likelihood. The order in which the observations are introduced mitigates the impact of outliers. We found that the SPL algorithm tends to find a substantial number of degenerate solutions when used to fit an FMM compared to a FMM fitted using the traditional EM algorithm. We propose a component self-paced learning algorithm that reduces the number of degenerate solutions. This algorithm uniquely determines the order in which observations are introduced at a component level for the estimation of the parameters of the FMM. The algorithm's performance is demonstrated using an extensive simulation study. It is observed that the algorithm reduces bias in the presence of outliers. An application of the algorithm on data from a real world scenario is also presented.

Mr AK Krishnannair (Department of Statistics University of Pretoria)

Co-author(s): Dr N Nakhaeirad (Department of Statistics University of Pretoria)

Predicting economic recessions using machine learning techniques

In the ever-evolving landscape of global economics, predicting and understanding economic recessions remain paramount challenges for policymakers, researchers, and financial analysts. The outbreak of the COVID-19 pandemic in 2019 has introduced unprecedented complexities, reshaping the economic dynamics of nations worldwide. This research gives the best performing models to assist businesses in predicting prior recession periods and identifies the most important variables to improve the overall performance of the models. To achieve this, in addition to artificial neural networks (ANN), machine learning techniques such as random forest and support vector machines are used to provide an efficient prediction model to avoid greater government deficits, growing inequality, significantly decreased income, and higher unemployment. Furthermore, an ensemble approach of ANN and Principal Component Analysis is proposed to be compared to the latter models. A real dataset on historical recession periods in African countries is employed to demonstrate the performance of the above-mentioned algorithms in practice.

Dr MA Lakeh (University of Pretoria)

Co-author(s): Dr N Nakhaei Rad (UP), Prof DG Chen (UP)

Pseudo-Observation Approach for Length-Biased Cox Regression Model

Pseudo-values are a way to estimate the expectation of a function of interest over the population when survival data is incomplete due to censoring or truncation.

Length-biased sampling is a special case of left-truncation when the truncation variable follows a uniform distribution. This phenomena is commonly encountered in various fields, such as survival analysis and epidemiology, where the event of interest is related to the length or duration of an underlying process. In length-biased sampling, the probability of observing a data point is higher for longer lengths or durations, leading to a biased sample.

The goal of this paper is to generate pseudo-observations to estimate the coefficients of a Cox regression model in the presence of length-biased right-censored data, in order to improve the accuracy and efficiency of statistical inference.

Mr K Mahloromela (University of Pretoria)

Co-author(s): Prof I Fabris-Rotelli (University of Pretoria)

Analyzing spatial point patterns on nonconvex domains

The analysis of spatial point pattern data is typically done to expand the basic understanding of the first- and second-order properties of the point process that generated the data. First- and second-order properties of spatial point patterns are estimated using density and distance-based measures. These measures rely implicitly or explicitly on the specification of the window domain. Thus, the correct specification of a window domain and the use of an appropriate distance metric to quantify proximity on the chosen window has an important role in the analysis of spatial point pattern data. Herein, we develop methodology to support the analysis of point pattern data in nonconvex window domains.

Dr MJC Malela (University of Pretoria)

Co-author(s): Prof FO Figueiredo (Universidade do Porto)

A Rank-Based EWMA TBEA Control Chart

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.

Prof SOM Manda (Department of Statistics, University of Pretoria)

Nonparametric Modelling for Survival Analysis

The proportional hazards model is widely used to model survival time in statistics. Useful extensions to the model include the stratified Cox proportional hazards model and frailty models that incorporate unobserved heterogeneity or random effects into the survival analysis. Some classes of nonparametric procedures for stratified hazards and frailty models are presented.

Ms NB Mashinini (University of Pretoria)

Co-author(s): Dr NN Rad (University of Pretoria), Dr J Nasejje (University of the Witwatersrand)

Investigating risk set size under different Cox regression model fitting techniques

In certain epidemiological studies, researchers investigate specific events, such as disease outcomes, and associated risk factors within a cohort. Analyzing the entire cohort can be time-consuming due to the vast amount of data. To address this, a nested case-control design can be employed, allowing for quicker and more efficient analysis by focusing on sampled cases and matched controls within the same population.

In survival analysis, the cohort dataset is crucial for defining risk sets in Cox model optimization. These risk sets are integral to the Cox partial likelihood function, which is used to fit the model. This paper applies the nested case-control design via a simulation study to these risk sets in the Cox partial likelihood function, exploring various case-control structures such as 1:1, 1:2, 1:4, and 1:8.

The study investigates whether the size of sampled risk sets impacts the time efficiency of the model and the precision of the estimated parameters using two optimization methods: Newton-Raphson and Stochastic Gradient Descent (SGD). Results from optimizing the four case-control structures using Newton-Raphson suggest that the Cox model estimates parameters converge to the true values more rapidly compared to using the full-risk set. Bias decreases with an increasing number of controls per case, with estimates converging to the truth after 1000 iterations with only 8 controls. When the Cox model is optimized using SGD with a complete risk set, it converges to the truth.

This study demonstrates how large datasets can be efficiently scaled and computation times reduced in survival analysis studies.

Ms AT Mberi (University of Pretoria)

Co-author(s): Prof S Manda (University of Pretoria)

Statistical Methods for Meta-Analysis of Individual Participant Data, with an Application to Women's Breast and Cervical Cancer Screening in Sub-Saharan Africa.

Traditional aggregate data meta-analysis relies on published results. However, individual participant data (IPD) meta-analysis has gained traction recently. IPD meta-analysis is advantageous as it provides an opportunity to investigate individual-level interactions. Two statistical methods for conducting an IPD meta-analysis are the one-stage and two-stage approaches. The one-stage approach involves pooling all individual participant data and performing statistical analyses simultaneously, for example, using a hierarchical regression model with random effects. The alternative is the two-stage approach, which involves conducting IPD within the studies independently and combining the summary measures obtained from the analyses. Using data from the sixteen Demographic and Health Surveys (DHS) across nine Sub-Saharan countries from surveys obtained between 2009 and 2022, we investigated both the one-stage and two-stage IPD meta-analysis to assess breast and cervical cancer screening uptake in the sub-Saharan region. Data on 127 317 women aged between 15 and 49 years was used to summarise the association between the level of education and type of residence with cancer screening uptake among women in sub-Saharan Africa. This study seeks to comprehensively understand the theoretical underpinnings of IPD meta-analysis approaches and their applicability to a critical regional public health concern.

Ms BLM McCall (University of Pretoria)

Co-author(s): Prof SM Millard (University of Pretoria), Prof FK Kanfer (University of Pretoria)

Investigations into Variants of Distributed Expectation-Maximisation Algorithms for Gaussian Mixture Modeling

Finite mixture models are widely used across industries and disciplines, with the Expectation Maximisation (EM) algorithm being a popular choice for parameter estimation. However, its inefficiency in traversing the complete dataset multiple times in the expectation step and the need for centralised data storage poses a challenge in the era of big data and distributed computing. In this presentation we explore alternative EM algorithm variants suitable for distributed environments, focusing on application of Gaussian mixture modeling. These distributed variants include incremental EM, parallel distributed EM, and asynchronous EM. The practical utility of these approaches amidst the landscape of big data analytics will be considered.

Mr PS Mdletshe (University of Pretoria)

Co-author(s): Dr R Thiede (University of Pretoria), Dr A Smit (University of Pretoria)

Markov chain accessibility model with traffic data

Accessibility modelling has been investigated since the early 60s. It is largely important in urban planning, freight business and policymaking. We hope to address the issue of accessibility using the South African road network within administrative units involving real life traffic data. This will be done through further enhancing the Markov chain accessibility model incorporating traffic data into it. We hope to draw more insights on whether traffic volume has a direct impact on ease of access, that being our measure of accessibility. The addition of traffic data can assist us in addressing several issues related to the South African road network like routes ambulances could take to avoid treacherous traffic conditions. Accessibility studies can truly add value in different aspects as it can be used as a factor in decision making that affect road networks in South Africa.

Dr N Nakhaeirad (Department of Statistics, University of Pretoria)

Co-author(s): Dr A Whata (Department of Statistics, University of Pretoria)

Dynamic joint-modeling of multivariate longitudinal and survival outcomes using deep learning

Joint modeling of the longitudinal and survival data enhances time-to-event predictions by including longitudinal outcome variables alongside baseline covariates. However, in practice, joint models can be restricted by parametric assumptions in both the longitudinal and survival sub-models. Additionally, computational challenges emerge when dealing with multiple longitudinal outcomes due to the numerous random effects that need to be integrated into the full likelihood. In this study, we adapt deep learning algorithms such as Dynamic-DeepHit, MFPCA-Cox, MFPCA-DeepSurv, MATCH-Net and Transformer-JM to determine the best-performing model in a multivariate joint modeling framework. Furthermore, we utilize post-processing statistics that provide clinical insight through measuring the influence of each covariate on risk predictions and the temporal importance of longitudinal measurements, thereby enabling us to identify covariates that are influential for different competing risks.

Ms J Nel (University of Pretoria)

Co-author(s): Mrs R Stander (University of Pretoria), Prof IN Fabris-Rotelli (University of Pretoria)

Bandwidth selection in a generic similarity test for spatial data when applied to unmarked spatial point patterns

The testing of similarity between spatial point pattern data sets is crucial for evaluating the quality and changes in spatial data. Different similarity tests have been developed for point patterns as well as images. A generic similarity test has been developed that handles any type of spatial data. When comparing unmarked point patterns using the generic similarity test, a kernel density estimate (KDE) is calculated for each point pattern. To calculate the KDE, a bandwidth value is required which is user-defined. The bandwidth is the smoothing parameter of the kernel. The focus of this work is to assess the effect the bandwidth choice has on this similarity test. The bandwidths considered in this work are specifically for spatial data sets. A simulation study is done to evaluate the effect of the different bandwidths on the similarity test. From the simulation study and application, it is seen that the similarity test could be sensitive towards the choice of bandwidth depending on the point pattern's characteristics such as homogeneous or inhomogeneous, the number of points being compared etc. We illustrate this similarity test and the effect of the different bandwidths on crime locations in Khayelitsha, Western Cape, South Africa.

Mr A Ngwira (Sokoine University of Agriculture)

Identification of common spatial and temporal trends in the epidemiology of cattle bovine tuberculosis and human extrapulmonary and drug-resistant tuberculosis in Malawi

Background: Identification of common spatial and temporal disease trends provides important information for the design of integrated disease control and monitoring programmes. We set out to identify the shared risk trends between cattle bovine tuberculosis (BTB) and human extrapulmonary (EPTB) and drug-resistant tuberculosis (DRTB) in Malawi.
Methods: A retrospective study of cattle BTB and human EPTB and DRTB cases from 2018 to 2022 was conducted. A shared component spatio-temporal model was fitted between BTB and EPTB and between BTB and DRTB, with district, year, cattle density, human density, temperature and precipitation as independent variables.
Results: The study has found significant positive spatial and temporal correlation of cattle BTB and human EPTB and DRTB. The shared risk patterns have a west-east gradient and individual disease risk trends have shown a south-north gradient. The predicted risk for all forms of TB in 2023 and 2024 has been decreasing and localized in the southern region. Precipitation seemed to be the common risk factor of cattle BTB (β: -0.012, 95% Confidence Interval (CI): -0.031, -0.002), and human EPTB (β: -0.002, 95% CI: -0.007, 0.000). Human population density was a risk factor of human EPTB (β: 0.005, 95% CI: 0.001, 0.008).
Conclusion: Integrated One Health interventions to reduce the burden of cattle BTB and human EPTB and DRTB may prioritize districts in the west and those in the southern region. The interventions may focus on minimizing changes in climate such as reduction in precipitation. Strategies to reduce EPTB in humans may target densely populated areas.

Mr AF Otto (Department of Statistics, University of Pretoria)

Co-author(s): Prof A Bekker (Department of Statistics, University of Pretoria), Prof JT Ferreira (Department of Statistics, University of Pretoria), Prof A Punzo (Department of Economics and Business, University of Catania), Dr SD Tomarchio (Department of Economics and Business, University of Catania)

A refreshing take on the inverted Dirichlet model via a mode-parameterisation for insightful multivariate clustering

In recent decades, there has been significant interest in exploring flexible and asymmetric probabilistic models, with a notable emphasis on the mode as a more intuitive measure of location compared to the mean or median. This study introduces and investigates a convenient mode-parameterised inverted Dirichlet model (also known as a Dirichlet type II/multivariate inverted beta distribution), which improves the implementation of the inverted Dirichlet across various statistical domains, particularly in nonparametric, model-based clustering, and robust statistics. Additionally, we illustrate the interpretability and impact of this model using real data within a clustering framework, to emphasise the value of the mode viewpoint.

Ms L Potgieter (University of Pretoria)

Co-author(s): Prof I Fabris-Rotelli (University of Pretoria), Dr R Thiede (University of Pretoria), Dr P Debba (University of Witwatersrand Johannesburg), Dr V Rautenbachb (University of Pretoria), Dr J Kemp (Stellenbosch University), Dr K Loggenberg (Stellenbosch University), Dr Z Munch (Stellenbosch University)

Enhancing Road Network Monitoring in Developing Countries

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 rural communities and promoting sustainable development.

Mr GC Singini (University of Malawi)

Co-author(s): Prof SOM Manda (University of Pretoria)

A Bayesian Markov Switching Model for Dynamic Prediction of Infectious diseases

Accurate infectious disease forecasting could support public health policy makers in designing appropriate preparatory and mitigating responses in preventing and reducing future infections. Most infectious diseases exhibit rising and falling progressions (states) over time and the widely used autoregressive integrated moving average (ARIMA) type models are unable to capture nonlinearity time trends in the observed disease occurrences that often depend on internal factors, which are not directly observable. Two of the most commonly used nonlinear time series models that can capture more complex dynamic patterns of infectious diseases by allowing switching between each states are the Markov Switching Models (MSM) and Hidden Markov Models (HMM). This study compares the forecasting performance of MSM and HMM within an interrupted time series study design, using simulation studies. Our results show that the MSM performed better on short-term prediction and running time. We also illustrate the two models with an application to example data.

Mr SB Skhosana (University of Pretoria)

Co-author(s): Prof S. M. Millard (University of Pretoria), Prof F. H. J. Kanfer (University of Pretoria)

Fitting a Gaussian mixture of non-parametric regressions using a mixture of Gaussian mixture models

Semi-parametric Gaussian mixtures of non-parametric regressions (SPGMNRs) are a flexible extension of Gaussian mixtures of linear regressions (GMLRs). The model assumes that the component regression functions (CRFs) are non-parametric functions of the covariate(s) whereas the component mixing proportions and variances are constants. The model cannot be reliably estimated using traditional methods. A local-likelihood approach for estimating the CRFs requires that we maximize a set of local-likelihood functions. Using the Expectation-Maximization (EM) algorithm to separately maximize each local-likelihood function may lead to label-switching. This is because the posterior probabilities calculated at the local E-step are not guaranteed to be aligned. The consequence of this label-switching is wiggly and non-smooth estimates of the CRFs as a result of incorrect component identification. In this presentation, we propose a novel approach to address label-switching and obtain improved model estimates. The proposed approach has two stages. In the first stage, we propose a model-based approach to address the label-switching problem. We first reformulate the SPGMNRs model as a mixture of Gaussian mixture model (GMM). We propose a modified EM-type algorithm to estimate the mixture of GMMs. Estimating the mixture of GMMs is equivalent in simultaneously maximizing the local-likelihood functions. In the second stage, we propose one-step backfitting estimates of the parametric and non-parametric terms. The effectiveness of the proposed approach is demonstrated on simulated data and real data analysis.

Mrs R Stander (University of Pretoria)

Co-author(s): Prof IN Fabris-Rotelli (University of Pretoria), Prof DG Chen (University of Pretoria)

A geostatistical approach for predicting hotspots in spatial lattice data

The identification of hotspots in spatial data has become an important part of spatial analysis. In some applications such as crime analysis and disease mapping the predication of location of a hotspot becomes important for local governments to implement preventative measures. In this work, we propose the use of geostatistical techniques (such as spatio-temporal Kriging) to predict observations at the next time step. A newly developed hotspot detection method is employed that uses the Discrete Pulse Transform (DPT) on spatial lattice data along with the multiscale Ht-index along with the Spatial Scan Statistic as a measure of saliency on the predicted observations.. The overall trend at each location will be classified into emerging hotspot classes.

Ms M Steyn (University of Pretoria)

Co-author(s): Prof IN Fabris-Rotelli (University of Pretoria), Ms R Stander (University of Pretoria)

Nonparametric assessment of spatial autocorrelation and variable importance in random forest and boosted tree models

In the presence of spatial autocorrelation - some unknown, most often non-linear, interdependent distribution underlying geographical space, the application of standard statistical modelling techniques is immediately invalidated due to non-normal and dependent errors. Moreover, the spatial interactions and spillovers between geographical areas give rise to an intricate task of variable selection and importance assessment.

Standard, parametric regression modelling techniques are performed based on some prior assumption of an appropriate model in order to estimate and test regression coefficients for the assumed model. This is most often not feasible in the presence of spatial autocorrelation where spatial patterns are influenced by various underlying interactions and spillovers of geographical conditions and natural movements across geographical areas, making it a complex, largely unknown, multi-dimensional factor difficult to conceptualise into traditional statistical models designed for lower dimensional data with minimal uncertainty about the underlying structure.

We assess the effectiveness of machine learning models for variable importance measurement and for modelling the unknown, complex and non-linear underlying spatially dependent distribution; such as Random Forest, Boosted Tree models and recently proposed spatial extensions. The model assessments are performed via nonparametric techniques, such as k-nearest neighbours and the Kolmogorov-Smirnov test.

Dr RN Thiede (University of Pretoria)

Co-author(s): Prof IN Fabris-Rotelli (University of Pretoria), Prof P Debba (CSIR), Prof CW Cleghorn (University of the Witwatersrand)

A test for the homogeneity of spatial linear networks

Spatial linear networks exhibit a variety of patterns. Like spatial point patterns, they can be homogeneous or heterogeneous. Although there exist a wide variety of tests for the homogeneity of point patterns, no statistical tests currently exist to quantify the homogeneity of spatial linear networks. This research provides a statistical methodology to test for homogeneity in spatial linear networks, using point pattern methods. The methodology approximates spatial linear networks by point patterns, obtained by taking the midpoint of each line. Existing tests for homogeneity of point patterns are then applied to the point pattern representations of the linear networks. The methodology is applied to test for homogeneity of formal and informal road networks in South Africa. This research is in line with UN SDGs 3, 9 and 10.

Dr PJ van Staden (Department of Statistics, University of Pretoria)

Co-author(s): Mr W Botha (Department of Statistics, University of Pretoria), Mr CC Geyer (Department of Statistics, University of Pretoria)

Adjustive rating systems in rugby union

With adjustive rating systems, the two competing teams in a match exchange rating points based on a comparison between the match result and the predicted match outcome. In this study we propose an adjustive rating system for rugby union based on exponential smoothing and compare this system with two existing rating methods: World Rugby's probit rating model and an ELO-based rating system. In contrast to the two existing methods, the proposed system in this study uses margin of victory directly in the calculation of the teams' rating points. Further, we consider factors influencing match outcomes, including home advantage and the relative strength of the competing teams. The three rating systems are applied to and compared for the 16 teams in the United Rugby Championship (URC).

Mrs L Venter (University of Pretoria)

Co-author(s): Prof D Chen (Arizona State University), Prof I Fabris-Rotelli (University of Pretoria)

Assurance in Diagnostic Accuracy Studies

In diagnostic test development, traditional sample size determination methods often relied on achieving desired power based on an assumed treatment effect. This approach did not account for the uncertainty in the true treatment effect, potentially misestimating the probability of demonstrating a positive outcome.

O'Hagan (2001) introduced the concept of assurance as the unconditional probability that a trial would yield a statistically significant result, averaged over the prior distribution of the treatment effect. This method improved the utility of clinical trials by simplifying computation for normal, binary, and gamma distributed data and was applicable to two-sided testing and equivalence trials. It argued that choosing sample size based on assurance provided a more reliable measure of a trial's success probability than power based on an assumed treatment effect.

Wilson and colleagues (Wilson, 2021) expanded the concept of assurance to diagnostic test development, particularly during the diagnostic accuracy study phase. They proposed a Bayesian approach for determining sample size, leveraging information from the analytical validity stage to calculate the required sample size based on assurance. Applied to a diagnostic test for VAP, the assurance-based approach reduced the necessary sample size compared to traditional methods, given relevant prior information.

Further exploration into the topic of assurance in diagnostic clinical trials, including the investigation of non-conjugate priors and the expansion of assurance concepts to studies involving longitudinal data, would significantly contribute to the field of diagnostic clinical trials. This research would advocate for the adoption of assurance as a standard measure in diagnostic trials.

Mr MW Waja (University of Witwatersrand)

Co-author(s): Ms OM Motlogeloa (University of Witwatersrand)

The need for robust research methodology when studying climate and health in developing countries

The biostatistical approaches employed in health research, along with the types of data used and the way the data is handled, are critical to obtaining scientifically valid information in response to a specific research question. As a result of the heightened availability of big data, and the ease of computing biostatistics through programs such as R, there has been an increase in studies that use statistical approaches to suggest that correlational relationships exist between variables, where this may not actually be true. This presentation relates to a Commentary paper that we submitted to the South African Journal of Science, in response to a paper published in Aids and Behavior on 20th February 2024 authored by Trickey et al. (2024) which argues that drought increases HIV (Human Immunodeficiency Virus) transmission in sub-Saharan Africa by presenting a positive correlation between drought, poverty, sexual behaviour, and HIV contraction. We argue that the methodology of the study is problematic, and does not consider key confounding factors, utilises data which is subject to biases, and utilizes inappropriate means to define drought. We highlight the ways in which inappropriate statistical approaches, along with a failure to engage with the data being collected and analysed, can impact the validity of research findings. It is of paramount importance that researchers employing a biostatistical approach in their research remain cognisant of how data is generated, what the data represents, and that the statistical approaches used should align with their specific research question in an effort to maintain scientific rigour.

WORKSHOPS

Workshop 1: Workshop on Symbolic Data Analysis

Professor Yaser Samadi: Associate Professor of Statistics, School of Mathematical and Statistical Sciences, Southern Illinois University Carbondale, USA

In our modern era of vast data availability, analyzing large datasets has become a routine challenge, given to advancements in technology. This workshop focuses on the analysis of symbolic data, which often arises when datasets are aggregated based on scientific criteria. Symbolic data includes various forms such as lists, intervals, and histograms, and finds applications across different fields including physical, medical, and social sciences. Additionally, certain datasets inherently possess symbolic values, such as species data, information with measurement uncertainties, financial data, and confidential data.

Symbolic data differs from classical data in that it represents distributions in multidimensional space rather than points. This workshop aims to shed light on the nature of symbolic data and their emergence in various contexts. By comparing methodologies for classical and symbolic data analysis, we highlight key differences through practical examples. Notably, we caution against the common practice of using classical surrogates, like aggregated means, which may lead to the loss of valuable information inherent in aggregated observation sets.

Workshop 2: Spatial modeling with INLA

Dr Janet van Niekerk: Research Scientist in Statistics, King Abdullah University of Science and Technology (KAUST), Saudi Arabia

Integrated Nested Laplace Approximations is a tool that provides accurate and efficient Bayesian inference of GAMM-type models. In this workshop I will show that most spatial models, being on a regular or irregular lattice, high resolution continuous data, or point processes can all be formulated as latent Gaussian models and thus a coherent methodology can be used to fit these models. The link between the Matern model and the weak solution of a specific SPDE is crucial for the approach in INLA and this will be briefly discussed. A new novel non-stationary Besag model for irregular lattice data will be presented. I will use the INLA package in R exclusively, and participants can work with me in the workshop if they have it installed beforehand.

Workshop 3: Mathematical modelling in team selection sports

Professor Gary Sharp: Associate Professor, Department of Statistics, Nelson Mandela University

The ability to quantify an individual's sport skill, provides analysts with the necessary quantitative tools to model the performances in a team setting. This workshop will present a mathematical modelling framework to select the optimal team or teams in scenarios where initially the best team seems intuitive, but rather quickly becomes more complex as the sporting code evolves. The examples used in this workshop originate in relay swimming, where a team consists of four swimmers, but are easily extended to other sporting codes. The focus in this workshop will be on formulating mathematical models and using available software to obtain optimal solutions.