SASA 2012

54TH ANNUAL CONFERENCE OF SASA

54TH ANNUAL CONFERENCE OF THE SOUTH AFRICAN STATISTICAL ASSOCIATION
















5 - 9 November 2012

Nelson Mandela Metropolitan University
Port Elizabeth, South Africa

PROGRAMME & ABSTRACTS

sas

Nelson Mandela Metropolitan University
for tomorrow

INTRODUCTION

The South African Statistical Association (SASA) and Nelson Mandela Metropolitan University (NMMU) Department of Statistics are proud to host the 54th annual SASA conference. The SASA conference will be held from the 5 of November to 9 of November 2012 in Port Elizabeth.

We welcome all to the Nelson Mandela Metropolitan University (NMMU) in the beautiful coastal city of Port Elizabeth in the Eastern Cape. NMMU is the largest higher institution in the Eastern and Southern Cape. With approximately 25 000 students enrolled across six different campuses. The SASA 2012 conference will be held on the NMMU Summerstrand South Campus. This campus is one of the five NMMU campus across the Eastern Cape and is situated on a nature reserve. The nature reserve conserves fynbos unique to the area and is home to wildlife including springbok, zebra and red hartebeest.

REGISTRATION

Registration for the conference with take place in the following venues, at the times specified.

Monday & Tuesday (5-6 November)	08h00 – 09h00	Madibaz*, Summerstrand South Campus
Wednesday (7 November)	07h30 – 12h00	Foyer, Auditorium, Summerstrand North Campus
Thursday (8 November)	07h45 – 09h00	Foyer, Building 5, Summerstrand South Campus
Friday (9 November)	08h50 – 09h30	Foyer, Building 5, Summerstrand South Campus

* “Madibaz” is indicated on the Summerstrand South Campus map as number 14 (Food Court).

All queries can be directed to the staff manning the registration desks.

PARKING

Dedicated parking areas will be made available to delegates. Follow the signs once you enter the NMMU campuses.

NAME TAGS

All delegate’s should wear their name tags at all times to gain access to the lecture halls, tea breaks, lunches and social functions.

TEA & MEALS

Teas and lunches will be served in the Madibaz café during the allocated sessions in the programme.

INTERNET

Wireless internet will be made available to the delegates in Building 5 during the course of the conference.

SESSION VENUES

All session venues used during the conference can be found on the maps on pages 59 and 60. Use the following table to identify the locations of the venues on the maps.

Venue	**Map Location**
Madibaz	Summerstrand South Campus, # 14 (Food Court) Lower Ground Level
Madibaz, Umvelani	In Madibaz Café, Upper Level.
Lab 4	Summerstrand South Campus, # 9 (Embizweni) 2nd Floor
Lab 5	Summerstrand South Campus, # 9 (Embizweni) 2nd Floor
Rendezvous Café	Summerstrand South Campus, # 87 (Goldfields North) Lower Ground Level
Building 5	Summerstrand South Campus, # 5 (Sanlam Lecture Halls) Ground Level
Auditorium	Summerstrand North Campus, #217 (Goldfields Auditorium)

MEETINGS AND SOCIAL FUNCTIONS

Meet and Greet

Monday 5 November at 17h30 for 18h00

Venue: Rendezvous Café

SASA Executive meeting

Tuesday 6 November at 17h15

Venue: Madibaz, Umvelani

Meet and Greet

Tuesday 6 November at 17h30 for 18h00

Venue: Rendezvous Café

ICCSSA Board Meeting

Wednesday 7 November at 16h15

Venue: Building 5, Room 0001

SASA AGM

Wednesday 7 November at 17h30

Venue: Building 5, Room 0007

Welcoming Function

Wednesday 7 November at 19h15 for 19h30

Venue: Humewood Golf Club

Young Statisticians' Breakfast

Thursday 8 November at 07h30

Venue: Blue Waters Café, Marine Drive, Hobie Beach

Gala Dinner

Thursday 8 November at 18h30 for 19h00

Venue: The Willows, Marine Drive

GUIDELINES TO SPEAKERS AND CHAIRPERSONS

Speakers

- Double check time and date of your presentation.
- Arrive at your venue at least 10 minutes before the start of your *session* (*not your presentation*) to ensure that all equipment is sufficient for your presentation. At this time all presentations should be loaded onto the computer in the venue. Each venue will have personnel that will assist you in loading the presentation before the start of the session.
- Report to the chairperson of the session prior to the start of the session.
- Keep to the time allocated for your presentation (strictly 15 minutes for your presentation and 5 minutes for questions). The chair of the session will warn you have 5 of your allocated 15 minutes remaining, and again when your time is up. Once chairperson has indicated the end of your presentation, you have to stop immediately.
- You are not allowed to move your time session to any other slot.
- Laser pointers and clickers will be available from the session assistants.

Chairpersons of sessions

- Keep to scheduled times.
- No changes are to be made to the programme. All presentations must start at the time indicated in the programme.
- Check the attendance of all speakers prior to the start of the session and ensure that all presentations have been loaded on the computer by the assistant.
- Open the session by welcoming the delegates and speakers and be sure to make the following announcements:
    - All cell phones should be switched off
    - State the title of the session
    - For each presentation, state the presenters name and the title of the presentation.
- Warn speakers 5 minutes before the end of the 15 minutes allocated to the presenters.
- Allow questions according to time (i.e. the presentation and all questions should not exceed 20 minutes).
- Thank all speakers and delegates at the end of the session
- Report problems and absent speakers to the assistant.

The above instructions are intended as a basic guideline for the sessions.

Please use your own initiative in the sessions to keep them running smoothly.

ABSTRACTS

Bayesian Hierarchical Spatiotemporal Models In Epidemiology: A Case Study Of Tuberculosis In Kenya

Thomas N. O. Achia[1] and Margaret Lillian Ndubi[2]
[1]School of Mathematics, Statistics and Compute Science, UKZN, Pietermaritzburg, South Africa
[2]School of Mathematics, University of Nairobi, Nairobi, Kenya

With the advent of modern computers, having improved data capture and immense data storage facilities, public health and epidemiological data sets have become readily available through online databases and in formats amenable to rich statistical data analysis. Many uses of these data sets, in disease mapping, are frequently constrained to a single time period. Data sets used in disease surveillance are, however, now available for time windows of several years. It is now possible, as a result, to consider the analysis of disease maps which have both spatial and temporal dimensions. The most common format for observations is counts of cases of disease within small areas that are available for a sequence of time periods.

In this paper, we analyse the spatiotemporal distribution of Tuberculosis (TB) morbidity in Kenya over a period of seven years (2002-2008) using two different models for space-time variation in TB risk. We explore the statistical properties of the spatiotemporal interactions model suggested by Bernardinelli et al. (1995) where temporal trend in risk may be different for different spatial locations and may even have spatial structure. We also explore the properties of the Nested model suggested by Waller et al. (1997), where the hierarchical conditionally autoregressive (CAR) model is applied to each time point separately. The data used was aggregated on a yearly basis from 2002 to 2008 and was obtained from the Kenya Ministry of Health's Division of Leprosy, Tuberculosis and Lung Diseases (DTLTD).

The Nested Hierarchical CAR model was found to fit the Kenya TB data better than the spatiotemporal interaction. The prediction maps produced show that high TB risk in areas known to have elevated HIV and poverty risk. This information can be used by the DTLTD for planning purposes, allocation of resources' and even dissemination of TB information. It can also be used to strengthen other strategies such as poverty eradication in Kenya.

Health Care Insurance Pricing

Franck Adekambi and Salha Mamane
School of Statistics & Actuarial Science, University of the Witwatersrand

Keywords: Alternating renewal process, Discounted aggregate amount of benefit, Moment, Health Care Insurance Pricing, Force of interest.

This paper uses an Alternating Renewal Process to model the lengths of the healthy and sickness periods. The first two moments of the discounted aggregate benefits paid out up to an arbitrary time *t* is then derived.

Students' Knowledge And Perception Of Quality Control Measures In Educational Services In Yaba College Of Technology, Lagos, Nigeria

Adeyemi Davidson Aromolaran
Yaba College of Technololgy, Lagos, Nigeria

Keywords: Quality assurance, Educational services, Tertiary institution, Quality awarness

Service quality in educational institutions has become subject of controversy among regulatory authorities and public galleries. This has resulted in various tertiary institutions putting in place quality assurance department to ensure that academic standards are built and sustained among Nigeria institutions. It therefore becomes necessary to test the students' knowledge and perceptions of the quality control

programme put in place to address problems confronting attainment of qualitative education service delivery in Yaba College of Technology, Lagos. This paper objectives of seek to know the level of students' awareness of quality control existence in the college, the effect of knowledge of quality control on academic activities of the students (vis-à-vis lecture attendance , lecturers' punctuality). A total of 800 validated questionnaires were administered on stratified random population of the students, 720 were returned, from which 646 were accepted and processed, representing approximately 90% response rate. It was further revealed that 52.9% of the respondents admitted knowledgeable of the existence of quality control unit in the college, 72.3% admitted not knowing the location of the quality control unit. 63.0% do not know what the activities of quality control units entail. Cross factor analysis using chi-square test statistic showed that there exist association between the students' level of education (class) and variables like knowledge of Quality Control existence in the college, knowledge of office location of quality control unit and admitting being displeased with quality of teaching in the college. Following the findings, a number of recommendations were made like improving campaign on awareness of quality control to the students from their first contact with the institution.

Forecasting South African Aggregate Retail Sales Using Univariate Linear, Nonlinear, And Nonparametric Time Series Models

Goodness C. Aye[1], Mehmet Balcilar[2] and Rangan Gupta[3]
[1]Department of Economics, University of Pretoria. aye_goody@yahoo.co.uk.[2]Department of Economics, Eastern Mediterranean University, Famagusta, Turkish Republic of Northern Cyprus, via Mersin 10, Turkey. mehmet@mbalcilar.net; [3]Corresponding author. Department of Economics, University of Pretoria, rangan.gupta@up.ac.za.

Keywords: Aggregate retail sales; Forecasting; Seasonality; Univariate Time Series Models

Aggregate retail sales data have strong seasonal, trend, and business cycle patterns, and like many macroeconomic time series forecasting such series, particularly during upswings and downswings, proves to be difficult. This article compares out of sample forecasting performance of univariate linear, nonlinear, and non-parametric models for forecasting monthly aggregate retail sales in South Africa. The time series data covers the period from 1970:1 to 2012:5 and out-of-sample forecast comparison is made over the period 1987:1-2012:5. Both seasonal models and models with seasonal dummy variables and trigonometric functions have been included among the 20 models compared in the article. The forecasting performances of the models are evaluated using uniform, right, left and right and left tail weighted Diebold-Mariano forecast comparison tests. Results obtained in the study indicate that the aggregate retail sales can be forecasted quite accurately. In general, models that adjusts for seasonality using seasonal dummy variables outperforms seasonal models. Although nonlinear and nonparametric time series models do not outperform the traditional linear models, such as the ARIMA, they are more successful in forecasting extreme values, which are located on the tails of the distribution.

The Description Of The Extended Bimatrix Variate Beta Type II Distribution

Andriëtte Bekker, Kotie Roux, Karien Adamski and Schalk Human
Department of Statistics, University of Pretoria

Keywords: extended bimatrix variate beta type II distribution, hypergeometric function of matrix argument, Wishart ratios, zonal polynomials

We named the distribution that originates from monitoring the process covariance structure of p attributes (where samples are independent, having been collected from a multivariate normal distribution with known mean and unknown covariance matrix) as the extended bimatrix variate beta type II. Two matrix variates that correspond to the two time periods, immediately after the change in the covariance structure took

place, will be considered. Therefore this proposed bimatrix variate distribution arises from different Wishart ratios. Some interesting characteristics of this bimatrix distribution (that originates from the statistical process control paradigm), the exact expressions for the probabilities of run-length and the distribution of the charting statistic are studied.

Biplots For Investigating Differential Living Conditions Between South African Racial Groups

Tsiresy Pierre Bernard and Sugnet Lubbe
Department of Statistical Sciences, University of Cape Town

Keywords: Inequality, Biplot, Principal Component Analysis, Multiple Correspondence Analysis, Sampling Weight.

During Apartheid, the level of socioeconomic inequality in South Africa has reached some of the world's highest levels. Inequality will be defined here as the overall differences in income and living conditions between the main different racial groups. As a multivariate analysis technique the Biplot offers the opportunity to have a graphical and concise overview of this multifaceted issue. The purpose is to visually assess the evolution of this matter after the end of apartheid by means of analysing and comparing data from different years. The data that were used were the October Household Surveys (OHS) data of 1996, 1998 and the General Household Surveys (GHS) data of 2002, 2006 and 2009. Using Principal Component Analysis (PCA) biplots it has been observed that the differences have not narrowed much overtime. However, this initial analysis had two culprits. First, the data were mostly categorical in nature which makes the use of PCA biplots inappropriate. Second, this analysis did not take account of sampling weights which are of extreme importance for nationwide survey data. These issues will be tackled by making use of Multiple Correspondence Analysis (MCA) biplots and by incorporating sampling weights in the analysis in order to make the solution more representative of the South African population.

Quantification of Estimation Instability and its Application to Threshold Selection in Extremes

Tom Berning
Department of Statistics and Actuarial Science, University of Stellenbosch

The main goal of this paper is to propose a measure which quantifies the instability of estimates over a range of chosen values of some other parameter. The measure is used to identify a region where the estimates are considered “stable”. Methods for identifying stable regions are then developed. These methods are applied to threshold selection in an extreme value analysis context, where the perturbed Pareto distribution is fitted to observed relative excesses. As a result a more accurate estimator of the extreme value index is obtained. As a further application the instability measure is employed in second order parameter estimation. The resulting second order estimator is then used to refine the threshold selection rule.

Maths4Stats – To Be Involved Or Not To Be Involved – That Was The Question

Rénette J Blignaut
University of the Western Cape

The Statistics and Population Studies Department at University of the Western Cape would like to share its experience of being part of the Maths4Stats training in the Western Cape. This presentation will focus on how the training sessions were planned and how the work was allocated to each of the participating stakeholders.

At the onset of the training, a questionnaire on background information was completed by each participating educator. During each training session prior knowledge related to the topic taught, was

collected by means of a short 5-minute questionnaire. At the end of each training session knowledge was again tested by a similar questionnaire to establish if some learning had taken place. The results show that knowledge improved during the training sessions, but that future interventions are still needed to ensure that Mathematics educators become well-informed and confident to teach topics such as data handling, probability and regression analysis.

With the knowledge gained from our first Maths4Stats experience, we share our views on proposed future involvement in this endeavour. The hard work was rewarded by appreciation and thankfulness from the educators as well as from the Western Cape Education Department.

Statistical Quality Control With Autocorrelated Data

Michelle Botes
Department of Statistics, University of Pretoria

Statistical process control (SPC) is used to monitor a set of observations over time to determine whether a process is in a state of statistical control. A process with known or unknown parameters values, is in-control when there isn't a shift in the parameter values and is considered out-of-control when the parameter/s shift or change to some different value. One common assumption in SPC is that the observations are independent and identically distributed (iid). However, in many cases the observations obtained from various processes are autocorrelated. This is due to the fact that in some cases the autocorrelation is inherent within the process which cannot be seen as an assignable cause that can be removed from the process. If a process exhibits autocorrelation and the traditional control charts for iid variables are applied, the false alarm rate can be inflated. A number of approaches have been considered for control charts with autocorrelated data. Charts which suitably adjust the control limits include the EWMAST chart, the ARMAST chart and the Modified Shewhart chart. Another approach is to transform the data into independent observations and apply traditional charts to the data. Examples of this approach include the residuals charts and charts on data separated by time and space, the so-called less frequent sampling. Once the data are independently distributed, non-parametric approaches can be applied to it like the sign or the singed rank statistic. This new approach is explored in this work. For the residuals chart the observations are grouped into large groups and the sign chart is applied, whereas for the less frequent sampling the values are grouped into smaller groups and the signed rank chart is applied. The results and the findings will be discussed.

A Hierarchical Bayesian Model: Combining The Generalized Gamma And Generalized t-Distributions

Stefan S Britz and Prof DJ de Waal
University of the Free State

The gamma family of distributions is often used to model heavy-tailed distributions. One example of a heavy-tailed distribution that is of great relevance to South Africa is the distribution of the inflows into the Gariep Dam, since it is the occurrence of large observations that leads to financial losses for ESKOM, as well as agricultural and domestic damage for downstream communities. In this presentation a hierarchical Bayesian model is fitted to the inflows, which are assumed to follow a generalized gamma distribution. A generalized t-distribution is used for the explanatory variable - the observed rainfall in Bloemfontein.

The model is then applied to find the probability of the dam exceeding 100% capacity on the basis of the predicted rainfall and also to determine the amount of water release required to limit the probability of spillage to the desired level.

Non-Linear Regression Models For The Characterisation Of The Early Bactericidal Activity (EBA) Of Tuberculosis Drugs

Divan Burger and Robert Schall
University of the Free State

The early bactericidal activity (EBA) of tuberculosis drugs is conventionally characterised by the rate of change, over a time interval of 3 to 14 days, in colony forming units (CFU) per millilitre counted in serial sputum samples from tuberculosis patients. More recently, time to sputum culture positivity (in short, time to positivity [TTP]) has been investigated as an alternative assay. Profiles over time of both CFU and TTP data have conventionally been modelled using linear, bi-linear or non-linear regression of logCFU or TTP against time. In this presentation, a range of non-linear models are fitted to CFU data over time (using the SAS® procedure NLIN). The EBA per patient is derived over Day 0 to Day 14 (EBACFU[0 14]) from fits of these non-linear models to the data of individual patients. Furthermore, non-linear mixed effects models (using the SAS® procedure NLMIXED) are fitted to the data of all patients from a given trial.

Detection And Down-Weighting Of Outliers In Non-Normal Clustered Data

Tinashe Chatora
University of Cape Town

This presentation introduces a variance shift outlier model (VSOM) for clustered count and binomial data. The VSOM will be used for the detection and down-weighting of outliers. This model considers potential outliers as observations with inflated variance and uses random effects to model the overdispersion associated with a single observation.

For clustered count data a VSOM is formulated assuming an underlying Poisson distribution for all observations, with the random effects following a gamma distribution. The variance of the ith observation relative to the rest of the data is determined by assigning a random effect to that particular observation and determining its shift in variance. The status of the ith observation as a potential outlier is indicated by the size of the associated shift in variance.

The VSOM for clustered binomial data is formulated assuming an underlying binomial distribution for all observations, with the random effects following a beta distribution. Similarly to count data it will be shown that the size of the associated shift in the variance of the ith observation will determine the status of that observation as a potential outlier.

For illustration purposes the VSOM will be applied to real datasets for both count and binomial data.

Challenges Of Large Class Teaching And Ideas To Facilitate A "Student-Centered" Learning Environment

Yoko Chhana
School of Statistics & Actuarial Science, University of the Witwatersrand

"Research in teaching and learning motivates for a "student-centered" or "learner-centered" style of teaching as opposed to the traditional didactic approaches. A student-centered classroom takes into account the needs of students as a group and as individuals with the outlook of encouraging them to improve their skills and actively participate in the learning process. It is often perceived that "student-centered" learning is a lot easier to implement in small classes as opposed to large classes. Lecturers at tertiary institutions are frequently confronted with teaching large classes, particularly at first-year level. Large classes are generally considered as a major obstacle in ensuring quality education. This paper reflects on the challenges related to teaching large classes and discusses some ideas on facilitating a "student-centered" learning environment for large classes."

Predicting Zimbabwe's Annual Rainfall Using Darwin Sea Level Pressure Index

Retius Chifurira and Delson Chikobvu
Department of Mathematical Statistics and Actuarial Science, University of the Free State

In this paper Zimbabwe's annual rainfall is predicted using the Darwin Sea Level Pressure Index for March of the previous year. The paper uses a predictive regression framework. This paper explores the influence of Darwin Sea Level Pressure Index (Darwin SLPI) on annual rainfall variability. Results from the study show that a high Darwin SLPI for March is associated with below normal rainfall for the coming year. While a low Darwin SLPI for March predicts above average annual rainfall for the coming year.

A Comparison Of Three Phase I Control Charts

Margarethe Coelho
Department of Statistics, University of Pretoria

Several Shewhart-type Phase I control charts for the location of a distribution have been proposed in the literature. Three of the popular ones are examined and compared in this study. These are the parametric Phase I Shewhart $\bar{X}$- chart, based on normally distributed data and two distribution-free (nonparametric) charts, one proposed by Jones-Farmer et al. (2009) based on the standardized subgroup mean rank and the other proposed by Graham et al. (2010) based on the subgroup precedence counts from the pooled median. An extensive simulation study was performed to find, in the case of the distribution-free control charts, the control limits and in the case of the $\bar{X}$- chart the chart design constant for different combinations of the number m of independent samples of size n, and for a specified nominal false alarm probability (FAP) of 0.05. The in-control performances, as well as the out-of-control performances for isolated and sustained shifts in the mean, were compared for a number of distributions of different shapes, also by means of simulation. It is seen that the rank-based chart, as well as the median chart performed similarly to the $\bar{X}$- chart for normally distributed data, but for heavy tailed or skewed data they both outperformed the $\bar{X}$- chart, with the rank-based chart showing the best results overall. The results make a strong case for using nonparametric Phase I charts in practice.

Comparison Of Test Estimators For A Location Parameter From A Normal Model Under BLINEX Loss With Special Focus On Preliminary Test Estimation

Judy Coetsee , Prof. A Bekker and Dr. S.Millard
Department of Statistics, University of Pretoria

The BLINEX function, introduced by Wen and Levy in 2001 (see Communications in Statistics -Theory and Methods, 30(1), 147-153), is both bounded and asymmetric, which allowed the same flexibility as the LINEX loss function but it also has the added advantage of being bounded. In this presentation the restricted maximum likelihood estimator (RMLE), the unrestricted maximum likelihood estimator (UMLE) and the preliminary test estimator for a location parameter from a normal distribution is considered under the BLINEX loss function. The risk functions for the preliminary test estimator, the RMLE and UMLE are derived under this loss function and the different risk structures are compared both analytically and computationally. In order to motivate why the BLINEX rather than LINEX loss function should be used, the risk for the different estimators under BLINEX loss are compared to the risk of the different estimators under LINEX loss and it is shown that the LINEX expected loss is higher than BLINEX expected loss. In addition, two feasible Bayes estimators are derived under BLINEX loss; and a feasible Bayes preliminary test estimator is defined and compared to the classical preliminary test estimator.

Multivariate Statistical Process Evaluation By Procrustus Analysis For Complex Chemical Processes

Roelof LJ Coetzer[1], Ruan F Rossouw[1] and Niel le Roux[2]
[1]Sasol Technology Research and Development;
[2]Department of Statistics and Actuarial Science, Stellenbosch University

The collection of high frequency online data is a necessity in most chemical processes today. However, chemical processes are inherently very complex due to the number variables of interest and the dynamic nature of the process. Coal gasification is one specific example of such a process. In order to better understand and interpret these complex systems in pursuit of continuous improvements and optimization, the data must be available in real-time, efficiently analyzed and summarized, and visualized for empirical and fundamental interpretations.

This paper presents the use of Orthogonal Procrustus Analysis to quantify the differences between reactors for a chemical process subject to multivariate responses. Furthermore, some extensions to Procrustus Analysis are discussed. Specifically, some ideas on multiple criteria are presented, including biplot axis predictivity and standard prediction error (SPE). The implementation of the developed methods and tools for online monitoring on the production facility is briefly discussed.

Aspects Of Multi-Label Classification

Ivona Contardo-Berning and Prof. Sarel Steel
Department of Statistics and Actuarial Science, Stellenbosch University

Single label classification is concerned with learning from a set of instances that are associated with a single label l from a set of disjoint labels L. If the number of labels exceeds two, the problem is referred to as a multiclass classification problem. Multi-label learning problems are concerned with learning from instances where each instance is associated with multiple labels. Multi-label classification is a problem that emerges from several applications such as textual classification, music categorization, protein function classification and the semantic classification of images.

An empirical study is conducted in order to compare several problem transformation methods using four benchmark data sets. Problem transformation methods are independent of algorithms and they transform the learning task into one or more single-label classification tasks. The single-label classification problems are typically solved with a single label classification approach and then the output is transformed back into a multi-label representation. The transformation methods are compared using example and label-based evaluation measures. Linear discriminant analysis and support vector machines were used as classification tools.

Combining Binary Classifiers To Improve Tree Species Discrimination At Leaf Level

Xolani Dastile[1], Gunther Jäger[2], Pravesh Debba[3], and Moses Cho[4]
[1]Anti-Corruption and Security, South African Revenue Services (SARS), xdastile@yahoo.com
[2]Department of Statistics, Applied University of Stralsund, gunther.jager@gmail.com
[3]Built Environment, Council for Scientific and Industrial Research (CSIR), pdebba@csir.co.za
[4]Natural Resource Environment (NRE), Council for Scientific and Industrial Research (CSIR), mcho@csir.co.za

This paper focuses on the discrimination of seven different savannah tree species at leaf level using hyperspectral data. The data is small in size, high-dimensional and shows large within-species variability combined with small between species variability which makes discrimination between the tree species (hereafter referred to as classes) challenging. We focus on two classification methods: K-nearest neighbour and feed-forward neural networks for the discrimination of the classes. For both methods, direct 7-class prediction results in high misclassification rates. We therefore construct binary classifiers for all

possible binary classification problems and combine them using Error Correcting Output Codes (ECOC) to form a 7-class predictor. ECOC with 1-nearest neighbour binary classifiers result in no improvement compared to a 1-nearest neighbour 7-class predictor whereas ECOC with neural networks binary classifiers improve accuracy by 10% compared to neural networks 7-class predictor, and error rates become acceptable.

Modelling Of Correlated Soil Animals Count Data

Legesse Kassa Debusho[1] and Gudeta W. Sileshi[2]
[1]Department of Statistics, University of Pretoria, Private Bag X20, Hatfield 0028, South Africa
[2]World Agroforestry Centre (ICRAF), Southern Africa Regional Programme, Chitedze Agricultural Research Station, P.O. Box 30798, Lilongwe, Malawi

Ecological studies naturally result in correlated data. Ignoring these correlations can result in biased estimation of ecological effects jeopardizing the integrity of the scientific inference. Mixed effects models are likely to appeal to ecologists for handling correlated data (e.g. Sileshi, 2008), however careful consideration must be given to the interpretation of the parameter estimates from generalized linear mixed effects models with non-identity link functions. The objective of this study was to compare the generalized estimating equations (GEE) under different correlation structures and suggest appropriate models to describe the relationship between soil animal counts and covariates. The GEE with independence, exchangeable and AR1 correlation structures were compared using count data set of ants from soils under the agroforestry systems in eastern Zambia. The GEE model with AR1 correlation structure gave a better description of the data than did the independence and exchangeable correlation structures.

Optimal Designs For Two-Colour Microarray Using Linear Mixed Models

Legesse K. Debusho and Dibaba B. Gemechu
Department of Statistics, University of Pretoria

In the microarray experiment, since each microarray is probed with two differential labelled cDNA samples, the array can be considered as a blocking factor. Basic microarray experimental design is therefore a block design with block size two. Each array in the microarray experiments has to be prepared separately and this could cause variation in the arrays. Therefore, Kerr and Churchill (2001) and Wolfinger et al (2001) have remarked that array effects should be taken as random. When array effects are random or if the gene-specific variance components are introduced for the random effects of arrays then a block design model setup will yield a linear mixed effects model. In this paper we will discuss optimal experimental designs for microarray experiments using the linear mixed models, which are the models that are appropriate describing certain correlated responses.

Cumulative Sum (CUSUM) Control Chart: Refinements And Enhancements

Frederik Coenrad Delport
Department of Statistics, University of Pretoria

Control charts are helpful in determining the special cause variations so that a timely action may be taken. While the Shewhart-type charts are the most widely known and used control charts in practice because of their simplicity and global performance, the cumulative sum (CUSUM) control charts are useful and sometimes more naturally appropriate in the process control environment in view of the sequential nature of data collection. These charts, typically based on the cumulative totals of a plotting statistic are known to be more efficient for detecting certain types of shifts in the process. In this paper we give an overview of the parametric CUSUM charts including the different enhancements and refinements that are available

in the literature, for example, the well-known fast initial response (FIR) feature, the variable sampling interval (VSI) method, the adaptive CUSUM (ACUSUM) and various discrete- and continuous CUSUM control charts (such as the Bernoulli based CUSUM, Poisson CUSUM and optimal Exponential CUSUM). Guidelines and recommendations are provided for the CUSUM chart's design parameters H (decision interval) and K (reference value). We also examine the impact of a choice of the reference value, *K*, on the performance of the chart in terms of the amount of the shift in the mean. The in- and out-of-control performance of the chart are studied through extensive simulations on the basis of the average run-length (*ARL*), the standard deviation of the run-length (*SDRL*), the median run-length (*MDRL*) and some percentiles of the run-length for a number of distributions. Results show that the proposed enhancements to the CUSUM control chart are considerably more efficient than the standard CUSUM control chart. A summary and some concluding remarks are given.

Measures Of Income Inequality

Rarang Phillemon Dikgale, MR Makwela and Prof A Tessera
Department of Statistics and Operations Research, University of Limpopo Author e-mail address: Phillemon.Dikgale@ul.ac.za, Abebe.Tessera@ul.ac.za, Mokowe.Makwela@ul.ac.za

Keywords: income inequality.

Income inequality has always been of major concern to economists, politicians, trade unions, policy makers and researchers in general. Inequality in general and income inequality in particular, can be measured in a number of ways. Inequality measures discussed in this paper include the range, GINI index, the Theil's Entropy and the Atkinson inequality. The paper focuses mainly on the comparative analysis of income inequality in South Africa and United States of America. The paper also investigates trends in income inequality, income inequality by population groups and gender. Income expenditure data from Statistics South Africa and the United States Census Bureau data will be used in this paper.

A Review Of Non-Standard Applications Of SPC Charts

Mandla Diko
Department of Statistics, University of Pretoria

Several myths and misunderstandings exist about statistical process control (SPC). One is that SPC is solely for process monitoring and improvement in the manufacturing domain. However, MacCarthy and Wasusri (2000) have shown that the application boundaries extend considerably beyond manufacturing. Taking a clue from this, our study reviews reported non-standard applications of SPC from 2000 to 2012. Non-standard applications are classified into six groups according to the domain to which control chart techniques have been applied. For each domain the nature of the application is described and analysed with respect to the control chart technique used, the purpose to which the control chart has been applied, the performance measures used, the units of analysis and the data sources. In addition we also mention the benefits, limitations, barriers and facilitating factors related to such use/implementation of SPC charts. We discuss some findings of our preliminary analysis. In particular, we uncovered two additional application domains that were missing in the review of MacCarthy and Wasusri (2000). These are the animal production applications and applications to personal everyday situations.

High Dimensional (Penalised) Discriminant Analysis

Mark Dowdeswell[1]; Tea Jashashvili[2,3]; Kristian Carlson[2,4]; Damiano Marchi[2]; Robert Nshimirimana[5]
[1]School of Statistics and Actuarial Science, University of the Witwatersrand; [2]Institute for Human Evolution, University of the Witwatersrand; [3]Department of Geology and Paleontology, Georgian National Museum, Georgia; [4]Department of Anthropology, Indiana University, USA; [5]NECSA, South African Nuclear Energy Corporation.
Mark.Dowdeswell@wits.ac.za; Tea.Jashashvili@wits.ac.za; Kristian.Carlson@wits.ac.za; Damiano.Marchi@wits.co.za; robert.nshimirimana@necsa.co.za

Linear discriminant analysis (LDA) is widely used for classification and dimension reduction. In situations where a continuous signal has been discretised, there may often be many more "variables" than observations (e.g. pixels in an image). Consequently, (reliable) estimation of the covariance matrices required for LDA-based classification of such signals becomes difficult. We discuss some aspects of a variant of LDA (due to Hastie, Buja and Tibshirani) called penalised discriminant analysis (PDA) which involves regularisation, similar in nature to that used in ridge regression, and which addresses the n<<p problem whilst also accounting for and indeed taking advantage of the underlying continuous nature of the original signal. We also contrast this approach with other (Fourier-based) techniques for dimension reduction in this scenario.

The origin of this statistical problem lies in the field of palaeoanthropology where the distribution of the cortical thickness of the shaft of first metatarsals may hold information on the bending and compressive stress patterns associated with different patterns of locomotion. It may be of interest, for example, to construct discrimination functions to classify extant hominoids (with potentially different patterns of locomotion) on the basis of cortical thickness distributions (effectively on cylinders). Such discrimination functions may subsequently be used to further investigate the characteristics of recovered fossils.

Herbaceous Biomass Prediction From Environmental And Remote Sensing Indicators

Nontembeko Dudeni-Tlhone[1]. Abel Ramoelo[2,3]. Pravesh Debba[1]. Moses Azong Cho[2]. Renaud Mathieu[2].
[1]Decision Support and Systems Analysis, Spatial Planning Support, Built Environment Unit, Council for Scientific and Industrial Research (CSIR), P.O.Box 395, Pretoria, 0001, South Africa
[2]Earth Observation Research Group, Natural Resource and the Environment Unit, Council for Scientific and Industrial Research (CSIR), P.O.Box 395, Pretoria, 0001, South Africa
[3]Faculty of Geoinformation Science and Earth Observation, University of Twente (UT-ITC), P.O.Box 217, Enschede, 7500 AE, The Netherlands

Feeding patterns and distribution of herbivores animals are known to be influenced by quality and quantity of forage such as grass. Modelling indicators of grass quality and biomass are critical in understanding such patterns and for decision makers such as park managers and farmers to efficiently plan and manage their rangelands. This study focused on predicting grass biomass using remote sensing and environmental variables. Since some of these variables were highly correlated, multivariate techniques such as partial least squares (PLS) and ridge regression were used to predict grass biomass in the Kruger National Park and the surrounding areas. The results indicated that both the environmental and remote sensing indicators had potential to predict grass biomass. Ridge regression showed better results since it explained about 41% of variation in the grass biomass, compared to the PLS model which explained approximately 33% variation.

The Discrete Pulse Transform For Images

IN Fabris-Rotelli
Department of Statistics, University of Pretoria, 0002, Pretoria, South Africa,
inger.fabris-rotelli@up.ac.za

The Discrete Pulse Transform (DPT) for images has a sound theoretical setting. We present this theory and its applications in image analysis.

Comparison Of Objective Priors For The Censored Rayleigh Model

Johannes Theodorus Ferreira
Department of Statistics, University of Pretoria

The Rayleigh distribution, which serves as a special case of the Weibull distribution, is known to have wide applications in survival analysis, reliability theory and communication engineering. The censored model is considered, such as to derive objective Bayesian estimators of the unknown parameter using different loss functions not previously considered for this model, whilst simultaneously considering different objective prior distributions. These estimators are compared by calculating the risk functions and comparing them under the simulated risk via a Monte Carlo simulation. Future endeavors are fleetingly mentioned with reference to possible subjective prior extensions and other possible loss functions.

Curable Shock Processes

Maxim Finkelstein
University of the Free State

In most conventional shock models, the events caused by an external shock are initiated at the moments of its occurrence. In this paper, we study *the new classes of shock models* when each shock from a nonhomogeneous Poisson processes can trigger a failure of a system not immediately, as in classical extreme shock models, but with delay of some random time. Moreover, we employ a new type of a curable shock model, where each failure can be cured (repaired) with certain probabilities. This type of shock processes with delays and the possibility of cure has not been considered in the literature before. We derive and analyze the corresponding survival and failure rate functions.

Panel Data Regression

Dr Lizelle Fletcher, Ms S Surovitskikh, Prof B Lubbe**
*Department of Statistics, *Department of Tourism Management, University of Pretoria*
lizelle.fletcher@up.ac.za

Conventionally, data are classified either as cross-sectional observations or as time series data. Panel data is essentially the combination of cross-sectional and time series data. Typically, multiple observations are recorded for each case in the sample, observed over T time periods, but with T too small to allow for a standard time series analysis; however, using cross-sectional analytic methods will probably violate the assumption of independence across observations. This type of data occurs frequently in biostatistics and in the economic environment, where its application allows the study of the dynamics of change.

The basic panel data regression model with a one-way error component will be explained and the four common options for panel data regression – pooled ordinary least squares, fixed effects least squares dummy variables, fixed effects within group and random effects – will be discussed, as well as general guidelines to decide between using a fixed effects or a random effects model.

The technique will be illustrated with an application in aviation where the aim was to estimate and statistically quantify the impact of the South African aviation policy, as measured by the four variants of the Air Liberation Index, on air passenger traffic flows in five key markets (intra-Africa; SADC; West Africa; East Africa and North Africa) over an eleven year time period.

Automatic Interaction Detection For Longitudinal Data

Jacky Galpin
School of Statistics and Actuarial Science, University of the Witwatersrand

Tree based methods for detecting variables predictive of a continuous or categorical dependent variable are available in several statistical packages, for the case of a single dependent variable. The case of multiple dependent variables has also received some attention in the literature, but little has appeared for the case of longitudinal data.

This paper outlines the usefulness of the methodology, and discusses some possibilities for applying this to longitudinal data.

Optimal Design For Two-Colour Cdna Microarray Experiment

Dibaba B. Gemechu and Legesse K. Debusho
Department of Statistics, University of Pretoria

Microarrays are powerful tools for detecting the expression levels of many thousands of genes simultaneously. They belong to the new genomics technologies which have important applications in the biological, agricultural and pharmaceutical sciences. The application of this technology is however, especially for two-colour cDNA(complementary deoxyribonucleic acid), poses some experimental design challenges like: Which mRNA (messenger ribonucleic acid) samples should be competitively hybridized together on the same slide?' and 'How many times should each slide be replicated?'. Furthermore, microarrays are expensive; thus carefully planning of these experiments is fundamental. The general aim of optimal design for such experiment is to maximize the precision with which effects that are of particular interest can be estimated. In this paper we will investigate the designs historically used for two-colour cDNA microarray experiment.

Hidden Markov Model Extensions For Animal Movement Models

Victoria Goodall[1,2]; Paul Fatti[1]; Norman Owen-Smith[1]
[1]University of the Witwatersrand, [2]South African Environmental Observation Network

Hidden Markov models are gaining in popularity in their application to animal movement modelling. They have been applied to a variety of datasets collected for a number of species including wolves, bison and caribou in Canada, tuna off the Australian coast, North Sea cod off the coast of the United Kingdom and sable, buffalo and zebra movements in South Africa and others. Some of the characteristics of animal movement data include the potential for missing data as a result of missed GPS locations, changes in the observation frequency or variable time lags between locations. Seasonal or daily cycles can also be present within the data. We investigate how hidden Markov models can be extended to fit models to data with these irregular time lags and the ways in which seasonality can be built into the models. The models are fitted to movement data from sable antelope, buffalo, zebra and lion collected in the Kruger National Park. Some of these datasets span more than two years' worth of observations. This region has a seasonal cycle with most of the rain falling during the summer months characterized by hot summers and warm dry winters. The movement of animals is strongly influenced by food and water availability which are dependent on the climatic conditions. This means that it is important to take the seasonal variability into consideration during the movement modelling process.

Nonparametric CUSUM And EWMA Control Charts For Monitoring Unknown Location Based On The Exceedance Statistic

Mrs. Marien.A. Graham
Department of Statistics, University of Pretoria

Standard control charts are often based on the assumption that the observations follow a specific parametric distribution, such as the normal. In many applications we do not have enough information to make this assumption and in such situations, development and application of control charts that do not depend on a particular distributional assumption is desirable. Nonparametric or distribution-free control charts can serve this wider purpose. Two nonparametric control charts, based on the exceedance statistics, are proposed for detecting a shift in the location parameter of a continuous distribution; the one being a cumulative sum (CUSUM)-type chart and the other being an exponentially weighted moving average (EWMA)-type chart. Advantages of the proposed charts include robustness to the violation of distributional assumptions, resistance to outliers and the fact that the exceedance statistics can save testing time and resources as they can be applied as soon as a certain order statistic of the reference sample is available. A comparison with a number of existing control charts, comprising of the traditional (normal theory) CUSUM and EWMA charts for subgroup averages and the nonparametric CUSUM and EWMA charts based on the Wilcoxon-Mann-Whitney statistics, is made. It is seen that the proposed charts perform well in many cases and thus can be a useful alternative chart in practice.

An Overview Of Image Segmentation Techniques

J-F Greeff and IN Fabris-Rotelli
Department of Statistics, University of Pretoria, 0002, Pretoria, South Africa,
inger.fabris-rotelli@up.ac.za

We present an overview of some image segmentation techniques, employed to extract regions of interest. Examples, with comparisons, are presented for Iterative Selection, Balanced Histogram, Otsu's method, Wellner algorithm, Integral Image algorithm, Gaussian mixtures and Iterated Conditional Modes.

Sampling Design And Analysis In The Cardiovascular Risk Household Survey

Nomonde Gwebushe , Carl Lombard, Nasheeta Peer, Naomi Levitt, Krisela Steyn, Estelle Lambert
Medical Research Council of South Africa, University of Cape Town

In 2008 a Cardiovascular Risk Household Survey was conducted in Cape Town in order to estimate the prevalence of diabetes and hypertension and to assess their association with psychosocial factors. This survey used a multistage stratified cluster sampling design and sample weights were used to account for the unequal selection probabilities and sample realisation. Different statistical methods were used to analyse different outcomes .For descriptive purposes the crude diabetes prevalence and age standardised diabetes prevalence for the standard WHO world population were calculated. To test the association between diabetes and the psychosocial factors a survey based multiple logistic regression was performed adjusting for the known risk factor for diabetes. Exploratory analysis informed to final formal model.

Designs for Field Trials with Unreplicated Treatments

Linda M. Haines
University of Cape Town

The problem of selecting varieties of a crop such as wheat from a large number of test lines has a long history and at the same time, with modern plant breeding technology, is very relevant today. The problem of selection based on maximizing yield is exacerbated by the fact that only a little seed for each of the test lines is usually available. To address this issue test lines are not generally replicated in the field, that is only one plot is assigned to each of the test lines, and information on their relative yields is gleaned by introducing controls or, if possible, by replicating certain of the test lines. In this paper designs for field trials which accommodate unreplicated treatments are reviewed and some interesting results relating to the specific class of augmented designs are presented.

Modelling Track Records Using Compound Distributions

Erin Hanly[1], David Friskin[2], and Gary Sharp[1]
[1]Department of Statistics, Nelson Mandela Metropolitan University, [2]RGT Smart

This paper models the progression of track world records using a compound distribution model found in actuarial statistics. Simulation is used to obtain short-term forecasts based on the fitted compound models. These forecasts show linear improvement, and so a modification to incorporate a levelling off of performance is demonstrated.

Generating Guidance On Public Preferences For The Location Of Wind Turbine Farms In The Eastern Cape

Jessica Hosking[1], Mario du Preez[2] and Gary Sharp[1]
Department of Statistics[1]and Economics Department[2], Nelson Mandela Metropolitan University

There is general consensus that South Africa should be generating more power, *inter alia*, through harnessing renewable energy, such as wind, but not with respect to the location of such generating projects. This paper describes a wind farm project proposed for development in the Kouga Local Municipality, reports local resident's preferences on its nature and applies choice modelling to analyse these preferences. Two respondent groups were surveyed, distinguished by socio economic status. Respondents were presented two different onshore wind energy development scenarios and a *status quo* option. The scenarios differed by the combination of four elements: the distance of the wind turbines from residential areas, the clustering of the turbines (job creation for the underprivileged respondent group), the number of turbines and a subsidy allocated to each household.

The paper finds that both respondent groups support South Africa's wind energy developments, but there were concerns about the impact of the wind farm on the environment and tourism. The affluent respondent group were found to be more sensitive to the distance between the wind turbines and residential areas, while the underprivileged group were found to be most influenced by changes to the prospects for employment created by such renewable energy projects.

Origin And Application Of A Noncentral Generalized Multivariate Beta Type II Distribution

Schalk Human, Karien Adamski, Andriette Bekker, Kotie Roux
Department of Statistics, University of Pretoria

The derivation of the run-length distribution (i.e. the waiting time until the first signal) when monitoring the unknown spread parameter while the known location parameter encountered a permanent upward or downward step shift is considered. More specifically, we focus on the scenario when the observations from each random sample are independent and identically distributed normal random variables so that the spread and location parameters are the variance and mean, respectively. It is shown that the solution to the run-length distribution involves a sequence of dependent random variables which are constructed from independent non-central chi-squared random variables. These dependent random variables are the key to understanding the performance of the control chart used to monitor the variance and are our main focus. For simplicity, the marginal (i.e. the univariate and bivariate) distributions and the joint (i.e. the trivariate) distribution of only the first three random variables following a change in the variance is considered. However, to generalize the results, a multivariate generalization is also proposed which can be used to calculate the entire run-length distribution. We also show how the noncentral generalized multivariate beta type II distribution is linked to other well-known multivariate distribution such as the Dirichlet.

Exact Confidence Intervals For The *P* Quantile Based On Order Statistics Of A Two-Parameter Weibull Distribution

Peter Iiyambo
University of the Free State

Many exact and approximate statistical hypothesis tests and confidence intervals (CIs) for the parameters and the *p* quantile of location-scale family of distributions are based on the maximum likelihood method. However, difficulties are often encountered during the estimation procedures. Furthermore, explicit expressions for the exact confidence intervals are difficult to derive and the maximum likelihood method may require intensive programming. This study compares by simulation exact rank-based confidence intervals for the *p* quantile of a two-parameter Weibull distribution, with exact confidence intervals based on the maximum likelihood method. Here, the generalized least squares method is applied to order statistics of standardized observations to obtain CIs for the *p* quantile.

Latent Class Analysis in Substance use Research

Esmè Jordaan [1], Petal Petersen [2]
[1] Biostatistics unit, MRC, [2] Alcohol and Drug Abuse Research Unit, MRC

Latent class analysis is an increasingly popular tool that researchers can use to identify latent groups in the population underlying a sample of responses to categorical observed variables. A latent class is a variable indicating underlying subgroups of individuals based on observed characteristics. Membership in the subgroup is said to be “latent” because membership in a class cannot be directly observed. Typically, LCA is carried out in an exploratory manner where here does not exist a strong a priori hypothesis regarding the number or nature of the latent classes underlying the data. In such a case, a researcher can fit several proposed models to the data with each differentiated by the number of latent classes, and compare the resulting fit indexed to determine which best corresponds to the observed data. A major contribution to applied work on substance use is the identification of latent classes characterized by particular patterns of substance use. For this presentation, we used data from a cross-sectional study among pregnant women attending Midwife Obstetric Units (MOUs) in greater Cape Town. A total of 684 pregnant women, of whom 418 admitted to lifetime use of alcohol, completed an interviewer administered

questionnaire, including questions on risky behaviour and patterns of use over the past 12 months. Covariates can also be included in the model to test whether they predict membership of the latent classes, without actually assigning individual member to a class. Conducting of confirmatory LCA is also possible, in which testing of specific hypothesis regarding some of the parameters is done.

Using The Markov Chain Monte Carlo Method To Make Inferences On Items Of Data Contaminated By Missing Values

Innocent Karangwa and Prof D Kotze
University of the Western Cape.
ikarangwa@uwc.ac.za; dkotze@uwc.ac.za

Keywords: Markov Chain Monte Carlo (MCMC) method, Listwise deletion, Bayesian analysis, Missing data, Metropolis-Hastings algorithm, Missing Completely At Random (MCAR) data.

The Markov Chain Monte Carlo (MCMC) is a method that is used to estimate parameters of interest under difficult conditions such as missing data or when underlying distributions do not fit the assumptions of Maximum Likelihood processes. The objective of this process is to find a probability distribution known as a posterior distribution in Bayesian analysis that can be used to estimate target parameters. In this paper, we only consider a case where data are contaminated with missing values and therefore need to be adequately handled using missing data techniques before making inferences on them. A review of the mathematics involved in MCMC methods as well as the link between this process and Bayesian statistics is provided. With an illustrative example, we also discuss one of the MCMC algorithms used by researchers, namely the Metropolis-Hastings. We further look at the mathematics of the MCMC in the presence of missing values in datasets. Using real data, we compare inferences made on a dataset with no missing values and inferences made on simulated data with different rates of missingness (5%, 10%, 15%, 20%, 25%, 30%, 35% and 40%) using the Listwise deletion and MCMC methods of handling missing data. The former technique discards subjects with missing values in the analysis and the latter augments observed data with generated missing values. We highlight the performance of these two methods under the assumption that data are missing completely at random (MCAR) at different generated rates of missingness.

Analysis Of Non-Food Household Expenditures Using Multivariate Structured Additive Regression Models

Lawrence N. Kazembe
Statistics Department, University of Namibia, Namibia

Analysis and determinants of household expenditures like housing, health, education, clothing and transport among others has been of interest to public policy analysts and decision makers, as it provides input towards welfare optimization and poverty reduction targeting. Many times, the observed expenditures are modelled separately, assuming independence of other expenditures ignoring the fact that these are correlated since they are derived from the same fixed income. This paper proposes a joint model to analyze non-food household expenditures in Namibia based on the 2009/2010 Namibian National Household Income and Expenditure Survey. We use a multivariate structural additive regression (MSTAR) model which simultaneously estimates fixed, nonlinear, and spatial effects in the model while adjusting for correlation in the data. Model inference is fully Bayesian, with diffuse priors assumed for the fixed effects, penalized second –order random walk priors for the nonlinear effects, intrinsic conditional autoregressive model fitted for the spatial effects, while exchangeable priors were used for the unstructured areal effects. The correlation in the data was handled by assuming Wishart priors for the residual terms. We fitted separate models for each response and compared the results with the joint

model. Results clearly indicated the advantages of joint model, with conservative results than those obtained under separate models.

Partial Least Squares (PLS) Variable Selection Using A Hybrid Particle Swarm Optimization Algorithm

Martin Philip Kidd and Martin Kidd
Stellenbosch University
mkidd@sun.ac.za

Particle Swarm Optimization (PSO) is a technique based on the principles of genetic algorithms which can be used to find optimal solutions to problems with large solution spaces. It has previously been successfully applied to variable selection in a regression setting.

In this presentation it is firstly shown how PSO can be applied for variable selection on Partial Least Squares (PLS), and how the method is extended to simultaneously search for the optimal number of PLS components as well as the optimal subset of variables.

Secondly an adapted version of the PSO algorithm (hybrid PSO) is discussed where each member of the population (outer algorithm) is used as input to another PSO algorithm (inner algorithm). The outer algorithm focuses on diversification while the inner algorithm focuses on intensification.

A central composite design (CCD) experiment was conducted to determine the optimal inner and outer population sizes and number of iterations for the inner algorithm. Time taken to find the optimal solution of a simulated dataset was used as outcome measure.

Finally results will be presented to show the advantage of using this hybrid algorithm above the standard PSO algorithm.

Multivariate Statistical Process Monitoring By Combining PLS And PCA

Gerhard Koekemoer and Roelof LJ Coetzer
Sasol Technology R&D

Multivariate statistical process monitoring (MSPM) is an efficient data-driven fault detection and diagnosis approach for complex industrial processes. Traditionally, unsupervised latent variable methods, specifically principal component analysis (PCA) have been employed with great success in the monitoring of industrial processes. Partial least squares regression (PLS) is a supervised latent variable method. In the current application we consider a projection surface for the process data (X), by means of PLS regression, which is appropriate to describe the production (Y) of an industrial plant using stable plant data. A PCA biplot based on the error space between predicted X values and the realised plant process data is then constructed. Robust control limits based on the PCA scores are provided for out-of-control detection and variable contribution to unexpected behaviour. We illustrate the methodology using data from a commercial production process.

Analyzing Financial Time Series With Varying Volatility And Extreme Clusters

Frans F. Koning and Prof. D. de Waal
University of the Free State

We are considering extremes in time series with clusters, applied in particular to financial time series. All extremes will occur over time, but they are often considered independent from other recent values. In our case we consider the possibility that one extreme value is dependent on the previous number of values, of which we consider the special case that it only depends on the very last value. Another complication is introduced where the volatility is not constant, but a stochastic process in itself. This creates clusters of

extremes over time, typical of financial data. Extremes inside clusters may also be very different from extremes outside clusters. The tail dependence is measured with some interesting simulations and examples of the GARCH(1,1,) model.

Asymptotic Normality For The Nonparametric Estimator Of The Quintile Share Ratio

Tchilabalo Abozou Kpanzou and Tertius de Wet
Department of Statistics and Actuarial Science, University of Stellenbosch

Keywords: Inequality Measure, Asymptotic distribution, Simulation Study, Confidence Intervals.

The quintile share ratio (QSR) is a recently introduced measure of income inequality, also forming part of the European Laeken indicators which cover four important dimensions of social inclusion (financial poverty, employment, health and education). In 2001, the European Council decided that income inequality in the European Union member states should be described using a number of indicators including the QSR. Not much is known though about the theoretical properties of its (traditional) nonparametric estimator. In this paper the estimator is defined and its asymptotic distribution theory provided. Using a simulation study, some finite sample properties of the limiting normal distribution are explored and reported on.

An Investigation And Historical Overview Of The G/M And M/G Queueing Processes

C Kraamwinkel and IN Fabris-Rotelli
Department of Statistics, University of Pretoria, 0002, Pretoria, South Africa,
inger.fabris-rotelli@up.ac.za

We present a historical and theoretical overview of the more complicated and less used G/M and M/G queuing processes, which allow for non-specific arrival and service time distributions. Such a model provides a more general setting for model fitting of real data for which the Markov property may not hold.

Aspects Of The *V*-Soft Minimal Hypersphere As An Outlier Detector For Multivariate Data

Morné M.C. Lamont
Department of Statistics and Actuarial Science, Stellenbosch University
mmcl@sun.ac.za

Keywords: Hypersphere; Kernel function; Support vectors

Data sets with multiple variables (multivariate data) are quite common in many areas of research. Many multivariate statistical techniques for the analysis of such data have been developed over the decades. Many of these techniques have been developed under the assumption that the data comes from a multivariate normal distribution. While many data sets follow a multivariate normal distribution, there are also plenty of cases where the data sets deviate from normality.

This paper is concerned with outlier detection and we consider a technique proposed by Tax and Duin (1999) which can be used to detect multiple outliers in multivariate data. The aim of the paper is to illustrate outlier detection for two scenarios: first where the data is normally distributed and then where the data is not normally distributed. The *v*-soft minimal hypersphere, a non-parametric technique, was used by Tax and Duin (1999) to obtain a support region (confidence region) and outlier detector for multivariate data. We explain this technique and discuss some aspects to take into consideration when it is applied.

Hawkes Processes And Their Financial Applications

Brendon M. Lapham and Iain L. MacDonald
Actuarial Science, University of Cape Town

The self-exciting point process, now more commonly known as the Hawkes process, is a model for a point process on the real line introduced many years ago by Alan Hawkes. The distinguishing feature of such a process is that the occurrence of an 'event' affects the intensity function at all subsequent times. Over the years such processes have been applied in neurophysiology and seismology in particular, but more recently significant financial applications have appeared, e.g. in the work of Paul Embrechts and his collaborators. We give an introduction to Hawkes processes and describe some applications to South African financial data. These applications include them modelling of extremal events in financial time series.

An Overview Of Stepped Wedge Designs

Kerry Leask
University of KwaZulu-Natal

Stepped wedge designs can be viewed as a combination of cluster randomized trials and crossover trials. In stepped wedge trials, the intervention is rolled out sequentially to clusters of participants. The design is useful for instances where the intervention can only be made available in batches and where it is believed that the intervention will be beneficial to all participants.

The sample size calculation for this design is presented and methods of analysing the data are discussed. Specifically, the analysis needs to take various sources of variability into account, as well as the fact that observations from the control and intervention groups are correlated. Methods of analysing the data arising from these trials include paired tests and random effects models

Willingness-To-Pay For Improved Fish Stock Size And Levels In The Sundays River Estuary, Eastern Cape, South Africa: A Choice Experiment

Debbie Lee
Department of Economics, Nelson Mandela Metropolitan University

The Sundays River fishery suffers from over-fishing and high retention rates of undersized fish. Management intervention is required to ensure the fishery‟s long-run sustainability. A choice experiment shows that the physical size of fish stocks is the most important predictor of choice and requires immediate intervention through increased license fees.

Biplots Constructed From Distance Matrices

Niel Le Roux[1]; Sugnet Lubbe[2] and John Gower[3]
[1]Department of Statistics and Actuarial Science, Stellenbosch University; [2]Department of Statistical Sciences, University of Cape Town; [3]Department of Mathematics and Statistics, The Open University, UK

Keywords: Analysis of distance, biplot, categorical canonical variate analysis, categorical variables, sum of squared distances.

Many different measures of similarity or dissimilarity between samples can be used for computing a distance matrix. Biplot theory allows the construction of biplots from any distance matrix. The type of biplot thus obtained depends on the distance measure used. For example: Pythagorean distance leads to the common principal component analysis (PCA) biplot; Mahalanobis distance results in a canonical variate analysis (CVA) biplot; chi-squared distance enables various types of biplot associated with (multiple) correspondence analysis. Of particular importance is the case when all variables are categorical with samples falling into K recognised groups. The proposed technique is termed categorical canonical variate

analysis (CatCVA) because it has similar characteristics to Rao's canonical variate analysis, especially its visual aspects. It allows group means to be exhibited in increasing numbers of dimensions, together with information on within-group sample variation. Variables are represented by category level points, a counterpart of numerically calibrated biplot axes used for quantitative variables. Mechanisms are provided for relating samples to their category levels, for giving convex regions to help predict categories, and for adding new samples. Computation is minimised by working in the K-dimensional space containing the group means. An analysis of distance table is derived for exhibiting the contributions between and within groups. This can be broken down further into contributions arising from different dimensions and sets of dimensions, especially the fitted and the remaining residual dimensions. The latter may be further subdivided into the dimensions holding the group means and the distances orthogonal to them. An R package for performing a CatCVA is introduced.

Models For The Analysis Of Competing Risks

Francesca Little
University of Cape Town

Standard survival analysis models focus on the analysis of the time to an event of interest in the presence of censoring. Censored observations are those that do not experience the event of interest during the period of observation for various reasons, including end-of-study and lost-to-follow-up. Survival models, like the Cox proportional hazards model, assume that the censoring mechanism is independent of the time to the event of interest.

Competing risks refer to the occurrence of other events that prevent the occurrence of the event of interest or that substantially change its probability of occurrence.

I review the methods and models for the analysis of time to event data in the presence of competing risks and biased censoring and illustrate their application to

1. an analysis of time to ART initiation in a cohort of patients co-infected with TB and HIV, and
2. an analysis of time to first opportunistic infection among a cohort of HIV+ infants.

Variable Selection And Binary Classification For Infrared Spectroscopy Data

Nelmarie Louw, Sarel Steel and Hélène Nieuwoudt
Stellenbosch University

Infrared spectroscopy data sets arise in many applied fields and analysing such data poses interesting statistical problems. The work in this paper was inspired by the following practical problem. Infrared spectroscopy data are available on grapes intended for export. In time a percentage of these grapes discolour. A problem of commercial importance is to predict discolouration based on spectral measurements made soon after harvesting. In essence we are confronted with a wide data set, with p spectral variables (each corresponding to a wavelength) observed for $np <$ sample units. The primary objectives are variable selection, i.e. identifying the spectral variables that distinguish well between the two groups, and constructing a classifier for accurate identification of the grapes which will discolour. An existing proposal by Rossi et al. (2007) in a regression context starts with B-spline compression of the spectral variables, resulting in $rp <$ new variables. Specification of the degrees of freedom of the spline model is an important question and we present a new proposal in this context. Standard variable selection techniques are applied to select a subset of the new variables.Importantly, each of the selected variables is associated with a contiguous range of the original spectral variables. Classification is performed using the selected spectral variables. We report the results of an empirical study in which this proposal is compared to other proposals in the literature, viz. partial least squares discriminant analysis and the fused lasso (cf. Tibshirani and Saunders, 2005).

On Systems With Gradual Repair

Zani Ludick
University of the Free State

In this talk, we are interested in the performance characteristics (e.g., generalized availability) of a repairable system with periods of operation and repair that form an alternating renewal process. During operation of the system, its performance is characterized by a continuous, decreasing function. When the system fails, we assume that it is not completely inoperable, but is still capable of some reduced level/quality of output. The initial level of the output after the failure is random and depends on the value of this function just prior to the failure. Specifically, it can be equal to 0. The repair action gradually increases and restores the performance quality of the system to its initial level at t=0. This type of repairable system has not been considered in reliability literature thus far. We obtain and analyze the expressions for generalized availability and related characteristics. A simple example is also presented.

Efficiency Of The MCMC With EM Algorithm As A Multiple Imputation Method

Huibrecht Elizabeth Maartens
School of Statistics & Actuarial Science, University of the Witwatersrand

In most areas of research, missing data have become a problem that analysts must address. Imputation methods such as multiple imputation (MI) are commonly used to deal with missing data. In this study the effectiveness of the MI method of Markov Chain Monte Carlo (MCMC) with the Expectation Maximization (EM) algorithm is assessed using simulated data as the complete baseline. Three missing data mechanisms are introduced for different percentages of missingness and their effect analysed through regression analysis. The goodness-of-fit of the regression models is assessed using the Akaike Information Criterion (AIC) and the coefficient of determination (R^2).

Numerical Maximization Of Likelihoods, E.G. As An Alternative To EM

Iain L. MacDonald
Actuarial Science, University of Cape Town

In the fitting by maximum likelihood of models involving latent variables or missing data, the EM algorithm is sometimes treated implicitly as the only method available, or the best available. In those cases where the likelihood can easily be evaluated, it is far from the truth to suppose that EM is the only possibility. I discuss here a selection of applications of direct numerical maximization (DNM) of likelihoods. In most of these applications there are latent variables or missing data, and both EM and DNM are available as means of finding maximum likelihood estimates.

In some simple models, for instance finite (independent) mixtures of Poisson distributions, DNM seems preferable to EM. And in the interactive exploration of more complicated models, DNM has the advantage that one can thereby avoid the detailed work of deriving, coding and checking the E and M steps required for EM. Furthermore, the need to perform a numerical maximization at the M step, which can arise when EM is used for more complicated models, implies that one is in any case doing one numerical optimization per EM iteration.

Frailty Modeling Of Recurrent Events

Jacob Majakwara
University of the Witwatersrand

The Cox-proportional hazards model has been studied deeply in literature and a lot of extensions have been made for modeling recurrent events. In generally, Cox's assumptions are not met in reality and specifically for recurrent events, it performs badly. Correlation caused by event dependence and heterogeneity are not accounted for in general, though some models do take into consideration heterogeneity. Frailty modeling has been proposed as the best alternative and especially the conditional frailty model. This model takes into account the correlation caused by event dependence as well as the heterogeneity into account. Exploration of this model is made in this presentation and is compared to variance-corrected models. Its advantages, pit falls and the possible way forward is presented. Simulation is done to show, in general, that frailty models are better than variance-corrected models in recurrent events.

Volatility Modelling Of Heteroskedastic Shares Using Generalised Pareto Distribution

Rhoda Makhwiting, Caston Sigauke, Maseka Lesaoana
Department of Statistics and Operations Research, University of Limpopo
rhoda.makhwiting1@gmail.com

*Keywords***:** Extreme quantiles, EGARCH, Extreme Value Theory, Generalised Pareto Distribution.

The paper proposes a modelling framework that accommodates autoregressive moving average (ARMA) as well as leverage effect via exponential generalized autoregressive conditional heteroskedastic (EGARCH) and modelling the tail behaviour of the return distribution using the Generalised Pareto Distribution (GPD). We refer to this model as the ARMA-EGARCH-GPD model. The quantile function of the GPD is then used to estimate extreme quantiles. A Pareto quantile plot is used to determine the threshold. Daily data of the all share index at the Johannesburg Stock Exchange (JSE) over the period 2002 to 2011 is used. Empirical results show that the ARMA-EGARCH-GPD model yields more accurate estimates of extreme returns than the ARMA-EGARCH model.

A Comparison Of The Minimum And The Precedence Charts For Subgroup Data

Jean-Claude Malela-Majika
Department of Statistics, University of Pretoria

Development of control charts that do not require a particular distributional assumption is desirable in practice. Nonparametric or distribution-free control charts can serve this purpose. Chakraborti, Van der Laan and Van de Wiel (2004) proposed a class of nonparametric Shewhart-type control charts, called the precedence charts, using some order statistic of a Phase II sample as the charting statistic and control limits constructed from a Phase I reference sample. Albers and Kallenberg (2008) proposed a similar nonparametric Shewhart-type control chart where the plotting statistic is the minimum of a Phase II sample. In this paper we compare the minimum chart with a precedence chart. Theoretical properties are studied and the in- and out-of-control performances of the charts are examined through simulations. A summary and some concluding remarks are given.

Comparative Analysis Of Price Indices Of Electrical Appliances In South Africa

Happy Maluleke, Maseka Lesaoana, Rahab Makwela
University of Limpopo
maluleke.happy@yahoo.com, Maseka.Lesaoana@ul.ac.za, Mokowe.Makwela@ul.ac.za

Keywords: Time series analysis, ARIMA, exponential smoothing, neural network, price indices.

An analysis of price indices of electrical appliances in South Africa is performed using monthly data from Statistics South Africa for the period January 1998 to December 2010. Time series analysis (exponential smoothing) and neural network are employed in developing forecasting models. The results for single, double and triple exponential smoothing are compared and triple exponential smoothing is found to be the best model amongst the three to forecast the electrical price indices in South Africa. Comparing neural network, ARIMA and triple exponential smoothing results, neural network is found to be the best model for forecasting.

Multivariate Spatial – Tempora Modeling of HIV and TB Co-Infections in South Africa

Annah Managa
Department of Statistics, University of South Africa
managma@unisa.ac.za

Keywords: Interconnection, risk factors, joint modelling, multiple modelling.

Prevention and control of tuberculosis (TB), HIV and other sexually transmitted diseases (STDs), present complex public health challenges. Understanding the interconnection between TB, HIV and STDs epidemics is crucial in guiding the best intervention and control strategies that are cost effective. These types of diseases are known to share similar risk associated with lower socio-economic status and sexual behaviour.

This paper uses multivariate discrete distribution to model the three diseases outcomes. We also control for both subject specific factors, both observed and unobserved. In this way, we will be able to determine the relative importance of the modelled factors. Estimating the magnitude of co-infections, helps to inform health policies on effective prevention and control measures.

Smoothed Temporal Atlases Of Age-Gender And Cause Specific Mortality In South Africa.

Samuel OM Manda
Biostatistics Unit, South African Medical Council, Pretoria

Most mortality maps in South Africa and in many parts of the sub-Saharan region are static, showing aggregated count data over the years or at specific years. The inability in these mortality atlases to account for both spatial and time dynamic modelling limits the use and application of the vital statistics data for robust policies and programmes in the public health.

In this study, mortality from specific and all causes in the nine pronvinces of South Africa and their evolvement over time from 1997 to 2009 are analysed using spatial-temporal hierarchical Bayesian ecological models. The analyses were separately done for each age-group and gender combination and for the main causes based on the South African National Burden of Diseases classification. For the former analyses, both the mortality rates and the mortality rate ratios (SMR) were estimated; the latter comparing to the overall South African mortality. Gender-specific and common gender as well as cause-specific and common cause geogrpahic-level risk components were modelled using shared component models.

The results show differential risks profiles between pronvinces, age-groups and gender. The time and pronvicial differences are also observed between cause-specific mortality. In conclusion, dynamic geographical and time distributions of age-gender and main cause specfic mortality contribute to a better understanding of the evolvement in the recent time of the burden of diseases in South Africa. This provides useful information for effective monitoring and evaluation of public health policies and programmes targetting mortality reduction across time and sub-populations in the country.

Multivariate Spacings, Generalized Quantiles and Level Sets Quantiles

David Mason
University of Delaware and North West University

I shall describe the notions of multivariate spacings, generalized quantiles and level set quantiles that I have investigated in my research career, along with their estimation. I shall also discuss the methods from classical and modern empirical process theory and as well the Poissonization techniques that are used to establish their asymptotic properties. Much of this work was done with Paul Deheuvels, John Einmahl, Wolfgang Polonik and Frits Ruymgaart.

Youth Of South Africa: Comparing Those In Education, Training And Employment With Those Who Are Socially Excluded

Michael Medupi
STATSSA

Statistics South Africa (Stats SA) and the Education Policy Unit (EPU) of Witwatersrand University are collaborating on studying the interface between education, training and work. As part of the programme, this paper compares the youth of the country (aged 15-34 years, divided into four equal age groups), who are socially excluded (they are not in education or training and they are not gainfully employed– also known as NEET youth) with those who are not-NEET (those who are in education or training, or gainfully employed).It uses the data of the General Household Survey of July 2010 for this comparison.

This analysis has led to some notable differences between NEET and non-NEETyouth groups.

- The age category 20-24 years contains the highest proportion of NEET youth.
- Compared to non-NEET youth, NEET youth are more likely to be African, and female.
- They are more likely to be less well educated.
- They are more likely to be living in informal settlements or in traditional rural areas.
- They are less likely to have access to services such as reticulated water, safe sanitation and electricity.
- NEET youth are more likely than non-NEET youth to go hungry.

The previously disadvantaged under apartheid continue to be disadvantaged.

Whilst overall, NEET youth tend to be African, a change may be occurring. In the age category 15-19 years, the largest proportion of NEET youth are coloured.

In future, further analysis will be undertaken using various Stats SA and other data sets to gain a better understanding of the links between education, training and work.

Examining Long-Run Relationships Between The BRICS Stock Market Indices To Identify Opportunities For Implementation Of Statistical Arbitrage Strategies

Brian Meki and D Kotze
University of the Western Cape.
bmeki@uwc.ac.za; dkotze@uwc.ac.za

Keywords: Cointegration, Rank, Statistical arbitrage, Mean reversion, Pairs trading, Unit roots, Stationarity, Vector Error Correction Model (VECM), Structural breaks, Spread, Threshold.

Emerging markets are generally known to be highly volatile. As a result, these markets are characterized by “high risk-high reward” investment opportunities. Armed with this knowledge, this research questions whether or not a selection of emerging markets may share any possible long-run relationships. The markets chosen for study are the BRICS. The conclusions of any shared long-run relationships between the BRICS are inferred from the economies’ major stock indexes. The paper uses the cointegration technique to investigate the long-run relationships with particular interest paid to the structural break which resulted in the recent 2007-2009 recession. The results of the study are stock price-based and they show that the BRICS markets are cointegrated only when the break is accounted for. In addition, the break dictates the VECM structure by proving to have significant influence on the cointegration rank. Further analysis shows that two pairs of stock indices from the BRICS are also cointegrated. This lays the necessary foundation for the implementation of the pairs trading strategy, a statistical arbitrage cointegration-based technique. The strategy proved successful and provided maximum trade returns of 8.64%.

One-way ANOVA with unequal variances

Henri Moolman
Department of Statistics, University of Zululand

When testing for the equality of means using ANOVA the assumption of equal error variances is usually made. In many situations this assumption is not plausible. This presentation addresses the following two issues.

1. What are the consequences of performing an ANOVA with an assumption of equal error variances when, in fact, the variances are not equal?
2. How can the testing procedure be modified to cater for cases when the error variances are not equal?

Accounting For Covariance Structure Changes Over Time In Multivariate Statistical Process Monitoring

André G Mostert and Gerhard Koekemoer
Sasol Technology Research and Development

Multivariate statistical process monitoring (MSPM) by principal component analysis (PCA) is commonly used in process monitoring on data that is time dependent. In this presentation we investigate the value of accounting for a covariance structure that can change over time. The proposed model consists of two components namely a PCA component that describes the multivariate deviation about the general sample mean, and a time weighted PCA component that accounts for deviations about the weighted mean of the usual PCA error. The time weighted averages are calculated using a biweight kernel**.** Mahalonobis distances are then used to detect and interpret out-of-control behaviour as well as variable contribution to unexpected performance. The results are presented and compared to a usual PCA monitoring approach using data from a complex chemical process.

Multivariate Economic Statistical Process Control

Precious Mudavanhu and Pieta van Deventer
Stellenbosch University

Keywords: Phase II; Assignable Causes; Multivariate Quality Loss; Type I and Type II Errors

The problem of minimizing loss in the case of univariate statistical process control, taking into account some statistical principles has been studied in detail. However, attempts at minimizing cost in statistical process control, adhering to some statistical criteria in the multivariate case present some severe stumbling blocks. Using Hotelling's T^2 the determination of the probability of the type II error is a problem. This in itself is worrying besides the fact it prevents one to calculate the average run length – one of the most used criteria in SPC. Several researchers tried various approaches to find a unified procedure, each in his/her own way. We present a summary of the most basic approaches up to now. A specific approach by Chou, Liu, Chen and Huang adapted from approaches by Montgomery and Klatt as well as Kapur and Cho, is high-lighted as an example of a typical approach. The reason for this technique lies in the provision for the the type II error as well as the fact that provision is made for assignable causes in the mean vector as well as the covariance matrix.

Meta-analysis of Covariate Effects Reported at Multiple Time Points

Alfred Musekiwa, Prof. Samuel OM Manda and Prof. Henry Mwambi
Wits Reproductive Health & HIV Institute

Keywords: meta-analysis, multivariate, treatment effects, multiple points

Meta-analysis is the statistical combination of results from two or more similar studies. Meta-analysis of a single treatment or covariate effect such as the mean difference (MD), risk ratio (RR) or odds ratio (OR) at one time point is well-known. There are, however, some situations that require the meta-analysis of more than one outcome at a time. Such an example is the study of HIV/ AIDS biomarkers namely CD4 count and viral load especially in response to antiretroviral therapy. Separate independent meta-analyses of these two outcomes ignore the within-study correlation of these measures. Such a problem is resolved by the use of multivariate meta-analysis, which is also well known. There is a special case of multivariate meta-analysis where measurements of one outcome are taken at more than one pre-determined time points in a study. Treatment or covariate effects can then be calculated at each time point in a study. Meta-analysis of such treatment or covariate effects requires taking account of the within-study correlations of the outcomes. This study demonstrates the adaptation of multivariate meta-analysis to the meta-analysis of treatment or covariate effects reported at multiple points. A real data set will be used as an example.

Forecastability, Stability and Cost of Forecast Error

Senzo Myeni
Eskom

In this paper we challenge some of the expectations both from Management and Forecasting Analysts in the business world, in particular the understanding of the concepts of Forecast Accuracy definition, Error Measures, Forecastability, Stability of the Data Generating Process and Cost of Forecast Error. We further define Forecastability in the manner free of error measures and forecasting methodology and challenges still needing attention. We then move on to the concept of Stability of the data generating process looking at the use of Approximate Entropy as the measure thereof in relation to Forecastability. Lastly we look at an example of a Cost of Forecast Error in Inventory Planning and Production Scheduling environment and the Political Calculus involved in the determination of appropriate service levels.

Multinomial Logistic Regression to Predict a Merchant's Experience of Card Fraud Incidents

Mark Nasila[1] and Gary Sharp[2]
[1]Retail Banking, First National Bank; [2]Department of Statistics, Nelson Mandela Metropolitan University

Keywords: Card Fraud, Multinomial Logistic Regression, Backward Elimination, Cross Validation.

The economic situation being experienced globally has highlighted that the already-fragile global banking systems are subject to greater risk and acts of fraud. There are new challenges in dealing with fraud arising from a fast-changing information technology environment, where the internet, point of sale and automated teller machines have become some of the most important retail sector transacting channels. This research uses a multinomial logistic regression model to predict the likelihood of a merchant experiencing a card fraud incident after customers have made purchases using their card payment facilities. These predictions are based on how merchants perceive specific aspects pertaining to card fraud. These aspects were taken from a survey done on merchants in the Johannesburg Metropolitan Municipality. Significant aspects were selected by a backward elimination process. Significance testing on the logistic coefficients using Wald test and likelihood ratio identified the aspects important for determining whether a merchant is likely to experience card fraud incidents. Cross validation of the fitted model was done using a classification matrix which showed that at least 72% of the merchants were correctly classified as to whether they had experienced card fraud incidents in the past.

It Rained But I Have No Rainfall Data!

Nothabo Ndebele
School of Statistics & Actuarial Science, University of the Witwatersrand

The problem of missing rainfall data within datasets is a frequent concern in climate studies. It is therefore important to apply appropriate statistical methods for estimating these missing values. In such cases, various methods of imputation have been employed; either by utilising available rainfall data at the same weather station or data from those of neighbouring weather stations. A discussion on the merits and shortcomings of some of these approaches is given. The nature of the missing data affects the choice of the method of imputation. In particular, here we consider a data set where: a) the daily rainfall data are completely missing due to records not being taken on that day, and b) where daily data are missing but included as an accumulated reading on a subsequent day. An application of imputation on this data set is presented.

Application of Granger Causality to Energy Market Models in South Africa

Vanessa Ndlovu[1,2]; Igor Litvine[1]; Pierre Le Roux[1]; Vince Micali[2] and Oksana Ryabchenko[1]
[1]Nelson Mandela Metropolitan University, [2]Eskom
ndlovuvc@eskom.co.za

With electricity being an important component of economic development it is vital that the impact of the supply of electricity on the economic growth of the country be well understood. Currently not many studies have been done on the analysis of this relationship in South Africa specifically and how this relationship impacts specific sectors of the economy that contribute to the total GDP of the country.

This study will assume rigorous application of Granger technique with proper statistical verification of assumptions, selection of relevant variables and providing trusted statistical forecasts. Other causality techniques will be used to confirm the findings. The study will be supported with comprehensive dataset/database compiled from various sources (e.g. Statistics South Africa, Reserve Bank, World Bank, IEA, Eskom, etc.).

Reporting On Student Assessment of the Class Room Environment

Prof. Delia North
University of KwaZulu-Natal
northd@ukzn.ac.za

The aim of this talk is to add value to the monitoring and evaluation of class room practise at UKZN. Recommendations are made which aim to improve the current reporting system on class room practises by students as conducted by the Quality Promotion Unit at UKZN.

The Distribution of Data From Thorough QT Studies: -Normal or Log-Normal?

Wallina Oosthuizen and Robert Schall
Department of Mathematical Statistics and Actuarial Science, University of the Free Sate
oosthuizenw@ufs.ac.za

Thorough QT (TQT) studies usually have a repeated measures cross-over design. Data from such studies is conventionally assumed to be normally distributed and analysed using linear mixed models. Alternatively, such data can be analysed after logarithmic transformation, which would be appropriate if the data were log-normally distributed. In order to investigate the best scale for data analysis (untransformed or log-transformed) we initially performed normality tests on the data of a large group of TQT studies, both on the original and on the logarithmic scale. Furthermore, we used the bootstrap method to estimate the coverage probability and average length of confidence intervals for relevant treatment contrasts, both when data are analysed on the original and on the logarithmic scale, respectively.

Partial Least Squares Regression Biplots

O.V.F. Oyedele[1], S. Lubbe[1] and N.J. Le Roux[2]
[1]Department of Statistical Sciences, University of Cape Town;
[2]Department of Statistics and Actuarial Science, Stellenbosch University

Keywords: Partial least squares regression; PLS algorithms; Biplots; PLS biplot.

In science and technology one of the most common data analysis problems is the problem of how to model one or more response variables using a set of predictor variables. Examples include the modelling of the properties of chemical samples using their chemical composition in chemistry, the modelling of the quality and quantity of manufactured products using the conditions of the manufacturing process in quantitative structure activity relationship studies (Wold et al., 2001). Usually the modelling of the responses is done using the multiple linear regression (MLR), which works well as long as the predictors are fairly few and are poorly correlated to each other. However, with modern measuring instrumentation, such as spectrometers, chromatographs and sensor batteries, data can be very large, strongly correlated and sometimes incomplete. As a result, partial least squares (PLS) regression can be a useful tool for modelling especially when there is no practical need to limit the number of predictors.

Results found by most recognised statistical methods of analysis can be visualized graphically using some form of graphical display such as biplots - a joint graphical display of all rows and of all columns of a data matrix.

In this paper we will discuss PLS as well as the graphical displays of a PLS regression of a data set, which we have termed the PLS biplot. This biplot enables us to simultaneously represent the sample points and variables. Both the predictor variables and response variables are represented, as well as the matrix of PLS regression coefficients.

The Median Odds Ratio (MOR) for Interpreting Random Effects of a Multi-Level Logistic Regression Model

Trishanta Padayachee
Medical Research Council of South Africa

The median odds ratio (MOR) is an attractive alternative to the intra-class correlation coefficient (ICC) or coefficient of variation for the interpretation of cluster-level variation when fitting a multi-level logistic regression model. It is simply estimated as a function of the cluster-level random effect parameter. Logistic regression is appealing as it facilitates an odds ratio interpretation of fixed effect parameters. The MOR enables the quantification of cluster-level variation through the computation of odds ratios making the interpretation of cluster-level random effects directly comparable to the odds ratio interpretation of the fixed-effect parameters of a logistic regression model.

The usefulness of the median odds ratio (MOR) is demonstrated through the investigation of cluster-level variation on the effectiveness of a TB intervention.

Comparison Of Some Methods For The Identification Of Common Eigenvectors

Theo Pepler
Stellenbosch University

When comparing samples originating from several multivariate populations in an analysis, an assumption about the covariance structures of the populations is necessary. Between the two extremes of homogeneity or complete heterogeneity there is a range of options, one of which is the existence of common principal components (CPC). According to the CPC model, the population covariance matrices share the same eigenvector structure, but with different sets of eigenvalues.

With the range of available options, a general problem is deciding on how to select the most appropriate model for the covariance structures from the hierarchy of models between homogeneity and heterogeneity. Most of the previously proposed methods to identify common eigenvectors in several groups rely on parametric assumptions which are not valid in many practical applications.

In this presentation, the CPC and partial CPC models for two or more multivariate populations will be briefly explained. Several methods for identifying common eigenvectors in two groups will be discussed. A new non-parametric method for the identification of common eigenvectors, using bootstrap distributions, will be presented and the results from a simulation study comparing this to the existing methods will be given.

Examining The Link Between Unemployment And Inflation Using Johansen's Co-Integration Approach And Vector Error Correction Modelling

Sagaren Pillay
Statistics South Africa

In this paper bi-annual time series data on unemployment rates (from the Labour Force Survey) are expanded to quarterly rates and linked to quarterly unemployment rates (from the Quarterly Labour Force Survey). The resultant linked series and the consumer price index (CPI) series are examined using Johansen's co-integration approach and vector error correction modelling. The study finds that both the series are integrated of order one and are co- integrated. A statistically significant cointegrating relationship is found to exist between the time series of unemployment rates and the CPI. Given this significant relationship, the study models this relationship using Vector Error Correction Models (VECM), one with a restriction on the deterministic term and the other with no restriction.

A formal statistical confirmation of the existence of a unique linear and lagged relationship between inflation and unemployment for the period between September 2000 and June 2011 is presented. This relationship can be explored further for the development of appropriate forecasting models incorporating other study variables.

Comparing Predictions From Combined Nonparametric And Parametric Models For Estimating fCO_2 In The Southern Ocean

Wesley Pretorius[1,2] and Sonali Das[1]
[1]CSIR Built Environment, Pretoria; [2]Stellenbosch University
wpret@sun.ac.za, sdas@csir.co.za

Keywords: Nonparametric; Multiple linear regression; Prediction; Southern Ocean.

In this study we compare different weighted combinations of a nonparametric model and a parametric model to predict fCO_2 for the Southern Ocean using *in situ* data from SANAE49 Leg 6. The Southern Ocean is a complex environment which, in the past, has been under sampled with data availability being limited spatially as well as temporally. fCO_2 is only measurable *in situ*, while the primary predictors of fCO_2 are available both *in situ* as well as remotely from satellites. As such, prediction of oceanic fCO_2 is a challenge. Using *in situ* data, we first develop a nonparametric kernel regression model, and make use of multiple linear regression predictions where the 'curse of dimensionality' takes over. Thereafter, we compare twelve models, from the pure parametric model, through differently weighted combinations of the parametric and nonparametric model, and finally the pure nonparametric model. Error rates based on an unseen part of the data are used to compare the prediction performance of the eleven models. Results indicated that the parametric models presented a higher bias in prediction, while the nonparametric were exposed to the 'curse of dimensionality'.

Volatility Modelling of the Producer Price Index in South Africa

Selaelo Lynette Ramare and Caston Sigauke*
Department of Statistics and Operations Research, University of Limpopo
**Corresponding author: lynette.selaelo760@gmail.com*

Keywords: ARMA, GARCH, producer price index.

In this paper, we develop ARMA-GARCH type models for modelling volatility of the Producer Price Index (PPI) in South Africa using monthly data for the period 1982 to 2011. The results suggest that daily returns can be characterized by an ARMA (1,0) process. Empirical results show that ARMA (1,0)-GARCH(1,1) model achieves the most accurate volatility forecasts.

Rewriting, Re-Registering But Not Always Revising: Patterns Of Behaviour In ODL Statistics Service Module Students

Eeva Rapoo
UNISA

Statistics service modules are a challenge at the best of times, but UNISA's distance education tuition model poses further problems for both lecturers and students. Recent changes in UNISA's tuition systems have added to the challenges, but at least have given us an additional chance do research into teaching and learning of statistics in the distance education setting. In this talk, we share some of the lessons we have learned, and our interventions aimed at increasing the course throughput rates.

Comparing Two Poisson Means: A Bayesian Approach

Lizanne Raubenheimer[1] and Abrie van der Merwe[2]
[1]Department of Statistics, Rhodes University;
[2]Department of Mathematical Statistics and Actuarial Science, University of the Free State.
L.Raubenheimer@ru.ac.za

Keywords: Bayesian intervals, Poisson parameters, Power and size of test, Probability matching prior, Weighted Monte Carlo method.

The probability matching prior for two Poisson means is derived. A weighted Monte Carlo method is used when using the probability matching prior. The power and size of the test, using Bayesian methods, is compared to tests used by Krishnamoorthy and Thomson (2004). They compared the usual conditional test (C - test) to a test based on estimated p - values (E - test). Non-informative (objective) priors will be used for the Bayesian approach. The priors used for the Bayesian approach are the Jeffreys prior, the probability matching prior, the uniform prior and two other priors.

Investigating Success Rates Of First Level Statistics Students Before And After The Implementation Of An Online Homework System Called Aplia In 2012

Fransonet Reyneke and Lizelle Fletcher
Department of Statistics, University of Pretoria
fransonet.reyneke@up.ac.za

The University of Pretoria is faced with the continuing problem of low success rates of its first year Statistics students. In 2005, while investigating the matric mathematics prerequisite/entrance criteria for Statistics, it was established that the 2004 pass rate was a disturbingly low 46.53%. This bleak picture has not improved dramatically despite continuing efforts on the part of the Department of Statistics to address the problem. Pass rates during the past five years have remained in the order of 60% and this continues to be a matter of great concern. Statistics (as opposed to Mathematical Statistics) is compulsory for all students who enrol for a BCom degree at the University, many of whom are not proficient in mathematics. A huge problem facing the lecturers is the large proportion of students, approximately 25% every year, who are repeating the course. The two specific Statistics modules for BCom students, comprising a first-level course, have been designated as so-called *high impact modules* and are thus targeted for additional tutoring support, serving as impetus for the current research. Major changes in the mathematics matriculation curriculum in 2008 have compounded the challenges faced by educators to improve students' performance and enrich the learning experience of students.

Since 2006, various intervention strategies have been adopted by the Department of Statistics at the University of Pretoria to address the problem. This paper traces the effect of these mediations over the past couple of years, starting with a brief explanation of the set of problems unique to South Africa, followed by the impact of, firstly, compulsory homework assignments introduced in 2006. Using assignments as an aid to learning proved to be flawed and was hence replaced by a system of class tests in 2007. Lastly, an analysis is provided of students' performance before and after the introduction of the modified matric curriculum in 2008, to illustrate the impact of this change.

A rationale is provided for a more interactive intervention where students will be required to complete assignments online - with immediate feedback. This online homework system called Aplia was implemented in the first semester of 2012 for STK110. The impact of the system on the first semester success rates of the first level students will be discussed. The students' view on Aplia and the way forward with Aplia will be important indicators for 2013.

Spatial Variation In Under-Five Mortality In South Africa

Sulaiman Salau
University of the Witwatersrand
sulaiman.salau@wits.ac.za

This paper uses data collected from the 2007 Community Survey to investigate district level spatial variation in under-five mortality rate (U5MR) in South Africa. Exploratory Spatial Data Analysis (ESDA) techniques are used to examine the spatial structure and distribution of U5MR and spatial regression models employed to explore the relationships between U5MR and a set of predictors. Models are compared and the benefits of spatial data analysis in child mortality research discussed.

Diagnostic Tests For The Distribution Of Random Effects

Leonard Santana
Department of Mathematical Statistics, North West University

Fourier methods are proposed for testing the distribution of random effects in classical and robust multivariate mixed effects models. The test statistics involve estimation of the characteristic function of random effects. Theoretical and computational issues are addressed while Monte Carlo results show that the new procedures compare favourably with other methods.

Accuracy And Fairness Of Rain Rules For Interrupted One-Day Cricket Matches

Robert Schall and Dianne Weatherall
Department of Mathematical Statistics and Actuarial Science, University of the Free State

We investigate the relative merits of various rain rules for one-day cricket matches. We suggest that an interrupted one-day cricket match presents a missing data problem: the outcome of the complete match cannot be observed, and instead the outcome of the interrupted match, as determined at least in part by the rain rule in question, is observed. Viewing the outcome of the interrupted match as an imputation or prediction of the missing outcome of the complete match, standard characteristics to assess the performance of classification tests can be used to assess the relative merits of rain rules. In particular, we consider the overall and conditional accuracy and the predictive value of a rain rule. We propose two natural requirements for a “fair” rain rule, and show that a fair rain rule must satisfy an identity involving its conditional accuracies. Performance characteristics of rain rules can be estimated from a sample of complete one-day matches. Our results suggest that the Duckworth-Lewis method, currently adopted by the International Cricket Council, is essentially as accurate as and somewhat fairer than its best competitors. An alternative rain rule based on the iso-probability principle also performs well but might benefit from re-calibration using a representative data base that includes recent one-day matches.

Improved Probability Limits Design for Attribute Data

Sandile Shongwe
Department of Statistics, University of Pretoria

The *np* and the *c* charts are among the most popular control charts used in the industry for categorical data. Being based on the binomial and the Poisson distribution, respectively, there are questions about how close the false alarm rates (or the in control average run lengths (*ICARL*)) of these charts are to a given nominal value. In this study, we investigate finding improved probability limits for these charts when the process parameters are known (i.e. Case K). We also compare the run length characteristics of the charts based on the improved probability limits with those based on the commonly used Shewhart 3-sigma

limits and the conventional probability limits. Examples and empirical studies for a variety of process parameters and sample sizes indicate that the improved probability limits charts have better in-control performance. That is, the improved probability limits charts either perform the same, or outperform the Shewhart 3-sigma limits and the conventional probability limits charts in terms of the closeness of the attained false alarm rate ($AFAR$) and/or the $IC\ ARL$ to the target nominal values.

Tail Quantile Estimation Of Heteroskedastic Intraday Increases In Peak Electricity Demand

Caston Sigauke[1,], Andréhette Verster[2] and Delson Chikobvu[2]*
[1]Department of Statistics and Operations Research, University of Limpopo, South Africa;
[2]Department of Mathematical Statistics and Actuarial Science, University of the Free State, South Africa
**Corresponding author: csigauke@gmail.com*

Keywords: Conditional extreme value theory, daily electricity peak demand, volatility, tail quantiles.

Modelling of intraday increases in peak electricity demand using an autoregressive moving average-exponential generalized autoregressive conditional heteroskedastic–generalized single Pareto (ARMA-EGARCH-GSP) approach is discussed in this paper. The developed model is then used for extreme tail quantile estimation using daily peak electricity demand data from South Africa for the period 2000 to 2011. The advantage of this modelling approach lies in its ability to capture conditional heteroskedasticity in the data through the EGARCH framework, while at the same time estimating the extreme tail quantiles through the GSP modelling framework. Empirical results show that the ARMA-EGARCH-GSP model produces more accurate estimates of extreme tails than a pure ARMA-EGARCH model.

Modelling Daily Increases In Peak Electricity Demand Using The Generalized Pareto Distribution

Caston Sigauke[1], Delson Chikobvu[2], Andréhette Verster[2]
[1]Department of Statistics and Operations Research, University of Limpopo, South Africa
[2]Department of Mathematical Statistics and Actuarial Science, University of the Free State, South Africa

The Generalized Pareto Distribution is used to model extreme daily increases in peak electricity demand (PED). The model is fitted to South African data to make a comparative analysis with the Generalized Pareto-type distribution. PED is influenced by the “tails” of probability distributions as well as by means or averages.

Further Techniques In Estimating Temperature In An Area In Which Temperature Is No Longer Measured

Dr. Morné Rowan Sjölander
University of the Free State
sjolanderM@ufs.ac.za

Time Series Models for Paired Comparisons have successfully been introduced and applied to various situations in recent years. Specifically, we have used several of our Time Series Models for Paired Comparisons to estimate temperature in an area in which temperature is no longer measured, using the current temperatures of the surrounding towns, as well as historic temperatures of this town, and the historic temperatures of the surrounding towns. Models that we have used include the Linear Exponential Analytical Hierarchy Process Model, the Log Linear Exponential Analytical Hierarchy Process Model, the Linear Bradley-Terry Model and the Sinusoidal Bradley-Terry Model. We now look at further techniques, using these and other models, in estimating temperature in an area in which temperature is no longer measured, such as applying decomposition to the data, and taking into account how much the various other areas’ temperatures’ influences on the temperature of the area in question, in order do find better estimates of the missing data in terms of various criteria, such as mean squared error, mean absolute error and maximum absolute error.

Multivariate Nonparametrics In Practice

Prof Chris F Smit
University of Pretoria

There exists a large number of articles on multivariate distribution-free/ nonparametric methods in literature, starting since at least the 1950's and reaching a climax/peak with the text by Puri and Sen in 1971. Since the 1980's there was a renewed interest in the multivariate nonparametric field, starting with the seminal paper by Hannu Oja (1983) on *Descriptive Statistics for Multivariate Distributions.* The approach followed by many researchers like Puri and Sen in the 1960's (assigning signs and ranks component-wise to multivariate data observations) was found to be insufficient for multivariate data due to the fact that the procedures are dependent on the coordinate axes of the data (not invariant to linear transformations). This deficiency lead to the need to develop procedures that are invariant to linear transformation of the data and terms like *affine invariance* and *affine equivariance* became important concepts in multivariate nonparametrics.

Reading papers in this field from a practical point of view is by no means straightforward and easy. Apart from the need for the interested practitioner to acquaint himself/herself with the essential background, assumptions and terminology, the practitioner also requires user friendly computer programs in the field of multivariate analysis.

The focus of this paper is to approach a selected (but important) part of the literature from a practitioner's perspective in terms of understanding the terminology, assumptions and computing tools needed for some of the most basic techniques. A hierarchy of assumptions on the underlying multivariate distribution will be discussed, some important multivariate estimates of location and scatter (proposed in literature) will be introduced briefly and a number of multivariate nonparametric tests be discussed. The availability (or lack thereof) of computer programs will be highlighted.

Crime Mapping In Ekurhuleni

Elsabe Smit
School of Statistics & Actuarial Science, University of the Witwatersrand

Crime is a prominent issue in South Africa. The high crime rate has given the country a reputation as a crime capital. Although there has been a reduction in serious crimes over the last few years, the fight against crime is still a key priority of Government. An understanding of the distribution of crime in space can aid in developing crime prevention strategies. This paper describes how principal component analysis can be used to derive general and specific measures of reported contact crime. The distributions of these measures in the Ekurhuleni Metropolitan municipality are then mapped and evaluated for the period 2003 to 2011.

The Effect of Differential Skewness on the Factor Structure of Equal Interval Scale Data

Jaclyn Smith and Dr Lizelle Fletcher
Department of Statistics, University of Pretoria
lizelle.fletcher@up.ac.za

Factor analysis is a statistical technique that is widely used an applied in areas such as psychometrics, behavioural- and social sciences, marketing, product management and operations research. Invariably, the assumptions underlying this technique are violated. As early as 1941, Ferguson detected that differential skewness (differences in skewness levels of the variables) will influence the factor structure in that the variables will be grouped according to their skewness levels and not the content of the variables. In this paper, the effect of differential skewness on the factor structure of interval scale data when

conducting an exploratory factor analysis are explored by means of a simulation study. Various skewness differences or discrepancies were tested to determine if a safe margin of differential skewness can be established for which the statistics will still be satisfactory and precise. Different sample sizes were also assessed to determine the role sample size plays when the data are differentially skewed.

A linear model for valuating preferences of freshwater inflows into forty selected estuaries along the South African coastline

Melnick Smith, Johan Hugo and Gary Sharp
Department of Statistics, Nelson Mandela Metropolitan University

Estuaries are environmental and economic assets to the population. The health status of our local estuaries, however, is being compromised due to a steady decrease in the freshwater inflow and supply. Upstream abstraction for industrial and domestic use, for example, could lead to mouth closure.

To ensure proper water resource management, different water allocation costs and benefits need to be compared and analyzed to secure an optimum solution (Mlangeni, 2007). Like many environmental services yielded to man, estuary services are not traded in any markets. Alternative markets are thus sought to allow the estimation of the value of such services.

The involved benefits of water allocations are predicted in this study by use of the CVM which elicits respondents' willingness to pay (WTP) towards predetermined changes in freshwater inflows into estuaries. The CVM was applied throughout the Water Research Commission's (WRC) Project K5/1413 from 2000 to 2009 (Hosking *et al*., 2010, Main).

Data imputation was employed to create an imputed dataset, enabling the modelling of the public's WTP through regression techniques. The population's total willingness to pay (TWTP) was further estimated by aggregation. The primary aim of this study was to collectively analyze the compounded data provided by the WRC and compare the results with the findings of previous studies.

An Overview Of Noise In Signal Analysis

PK Smith and I N Fabris-Rotelli
Department of Statistics, University of Pretoria
inger.fabris-rotelli@up.ac.za

We present an overview of common methods for modelling a signal contaminated by noise. Specifically discussed are the structure imposed on a model, the properties the noise assumed, some common statistical distributions for the noise and the Signal-to-Noise ratio as an indicator of the quality of the contaminated signal.

Using P-values for Multiple Comparisons

Francois Steffens
Department of Statistics, University of Pretoria

In some application fields such as Biochemistry, researchers often observe very large numbers of variables on relatively few cases. The result is that most multivariate techniques involving the inversion of a covariance matrix are not applicable because the covariance matrix is singular. Doing analyses variable-by-variable is also problematic if adjustments for multiple comparisons, such as the Bonferroni adjustment, are made because each test becomes extremely conservative. The proposal is to look at the distribution of P-values, each of which has a uniform distribution under the null hypothesis, to find a clue as to the rejection of the null hypothesis or not.

Permutation Procedures for ANOVA, Regression and PCA

Christine Storm and Dr Lizelle Fletcher
Department of Statistics, University of Pretoria
lizelle.fletcher@up.ac.za

Parametric methods are effective and appropriate when data sets are obtained by well-defined random sampling procedures, the population distribution for responses is well-defined, the null sampling distributions of suitable test statistics do not depend on any unknown entity and well-defined likelihood models are provided for the nuisance parameters.

Permutation testing methods, on the other hand, are appropriate and unavoidable when distribution models for responses are not well specified, nonparametric or depend on too many nuisance parameters; when ancillary statistics in well-specified distributional models have a strong influence on inferential results or are confounded with other nuisance entities; when the sample sizes are less than the number of parameters and when data sets are obtained by ill-specified selection-bias procedures. In addition, permutation tests are useful not only when parametric tests are not possible, but also when more importance needs to be given to the observed data set, than to the population model, as is typical for example in biostatistics.

The different types of permutation methods for analysis of variance, multiple linear regression and principal component analysis are explored in this paper. More specifically, one-way, two-way and three-way ANOVA permutation strategies will be discussed. Approximate and exact permutation tests for the significance of one or more regression coefficients in a multiple linear regression model will be explained next, and lastly, the use of permutation tests used as a means to validate and confirm the results obtained from the exploratory PCA will be described.

The MDS-GUI: A Graphical User Interface For Comprehensive Multidimensional Scaling Applications

Andrew Timm and Sugnet Lubbe
Department of Statistical Sciences, University of Cape Town, South Africa

Keywords: MDS, GUI, tcltk

The MDS-GUI is an *R* based graphical user interface for performing Multidimensional Scaling (MDS) methods in a number of ways. The software was developed using the *R* wrapped tcltk package and a number of the packages affiliated it, such as tcltk2 and tkrplot. Upon completion, the software will be available to the public in the *R* package, MDSGUI. While the program is in its final stages of development, it is at this point fully demonstrable and is likely to be ready for submission to the CRAN database by the end of 2012. The intention of the MDS-GUI is that the menu structures and overall layout be set out in a way that is found to be user friendly and uncomplicated as well as comprehensive and effective.

The MDS-GUI has been developed to provide the user, even with no theoretical background on the subject, with the opportunity to perform a number of MDS methods and output a host of relevant details and graphics. In broad terms, the GUI allows the user to simply and efficiently input their desired data, choose the type of MDS they would like to perform as well as select the type of output they would like to achieve by the analysis. The use of sub-menus and property tabs gives the user the option to fine tune specific parameters of the desired MDS procedure as well as provide options to alter the way in which the resulting plots are displayed. The graphical outputs are of an interactive nature and allow the user to make adjustments to the output with a cursor to observe any difference in results. Multidimensional Scaling is usually an iterative technique, which is a quality preserved by the graphics of the software. The user is

thus able to have a visual display of the processes at work and observe the moving ordination configuration.

This paper will, first of all, discuss the MDS-GUI in terms of both its development and relevance from a Multidimensional Scaling point of view. This discussion will include an overview of the main features of the software as well the coding techniques and methods used in its construction. Following this will be a small case study which was analysed entirely using the MDS-GUI. The data in question used for the study is Rothkopf's (1957) famous *Morse-Code* confusion data. This data is well known in an MDS context and is thus used as means of confirming the validity of the results produced by the software.

Bayesian Inference For The Piecewise Exponential Model Using Objective Priors

A.J. van der Merwe and P.C.N. Groenewald
Department of Mathematical Statistics and Actuarial Science University of the Free State

An objective Bayesian approach for modelling the reliability of multiple repairable systems using the piecewise exponential (PEXP) model is considered. In particular, inference under different priors are discussed. The PEXP model assumes that the times between failures are independent and exponentially distributed, but the mean is allowed to either increase or decrease with each failure. It is one of the most popular and useful models in reliability and survival analysis. Reference and probability matching priors are derived for the mean failure time and predictive densities are given for future failure times. The application of the model to failure data on load-haul-dump machines given in Hamada et al (2008) is presented. Simulation studies to evaluate the performance of the proposed model are provided.

Adaptations Of KS Test With Applications

Sean van der Merwe and DJ de Waal
Department of Mathematical Statistics and Actuarial Science, University of the Free State

We show how to adapt the KS test to dramatically widen its possible areas of application. We consider two adaptations: using metrics such as the Mahalanobis measure to extend the application of the KS test to multiple dimensions in a natural way; and using the predictive posterior distribution to change the null hypothesis to a general test of whether a distribution is valid for a sample (as opposed to simply testing goodness-of-fit). We show varied example applications, including determining an optimal threshold in extreme value modelling and testing whether a sample follows a Dirichlet Type distribution.

Model Risk Management

Carel van der Merwe
Ernest and Young

Model risk poses an increasingly higher threat due to the ever more widespread use of models. In essence, mathematical models are used to try and capture an observed phenomenon. At first this risk might seem to be of little importance to corporate treasuries; however it could possibly be a significant risk that is sometimes overlooked.

We consider a proposed model risk management framework, and also present results on an international benchmarking survey on model governance and validation.

Multi-step Approaches To Consumer Segmentation

Marieta van der Rijst[1], Tormod Næs[2] and Nina Muller[3]
[1]Agricultural Research Council; [2]Nofima, Norway; [3]Stellenbosch University

In consumer research there are essentially three sources of information that need to be linked to each other: product descriptors, consumer preference of the products, and consumer background characteristics. The identification of relationships between these sources of information is the key to successful product development.

The methods for experimental consumer studies can broadly be split in two categories: preference mapping studies, which focus on intrinsic attributes, and conjoint studies, where the focus traditionally has been on extrinsic attributes. Several unified data analysis approaches, that incorporate both preference mapping and conjoint studies, have been suggested for improved insight into the main drivers of liking in different consumer segments. Some of these approaches consist of two (or more) steps, while other methods directly deal with the different sources of information. The paper presents a practical application of multi-step approaches to consumer segmentation.

South African Pension Funds: A Review Of Construction Strategies And Empirical Evidence

JD van Heerden
Stellenbosch University

In this article we investigate whether it is optimal for South African pension funds to allocate the full twenty-five percent of available assets to offshore asset classes as allowed under the revised Regulation 28. Asset allocation optimisation strategies differ with respect to the optimised objective, and we compare the results of seven commonly used asset allocation optimisation models over a 10-, 20- and 30-year investment period. The majority of optimisation models show that domestic-only funds significantly outperform funds with a foreign allocation component over the 10-year investment horizon. Over a 30-year period only one strategy recommends a domestic-only portfolio, three strategies recommend portfolios with a twenty-five percent foreign exposure that significantly outperform their domestic-only counterparts while the remaining three are indifferent between domestic-only or foreign-allocation portfolios. Applying a bootstrap approach we find that the re-sampled mean-variance optimisation strategy applied using a 20-year historical data period is the superior optimisation strategy when measured against four different benchmarks over the three investment horizons.

Bayesian Estimators Of The Location Parameter Of The Normal Distribution With Unknown Variance

Janet van Niekerk and Andriette Bekker
Department of Statistics, University of Pretoria

The estimation of a location parameter of the normal distribution has been widely discussed and applied in various situations. The Bayes estimators under Linear Exponential (LINEX) loss are functions of the moment generating function of the Student's t-distribution and are therefore unknown. In this paper the explicit Bayes estimators of the location parameter of the normal model for two different loss functions, the re1flected normal loss and the LINEX loss functions are proposed and evaluated. The performances of these estimators are evaluated using Monte Carlo simulation.

Statistical Techniques To Calculate Loss Given Default

Jaco van Tonder
Ernst and Young
jaco.vantonder@za.ey.com

The global accounting standards and the Basel regulations require banks to assess the assets (loans) on their books for impairment. This impairment figure impacts both the profits of the bank and the capital the bank must hold against loans. The impairment calculation is a combination of the exposure at default, probability of default and loss given default.

Loss given default (LGD) is a particularly important component of the bank's impairment calculation as it is very sensitive to the assumptions made and to wider economic circumstance. However, despite the importance of this component and the fact that banks are directed by one set of regulations, bank follow very different approaches to LGD calculations. The aim of this presentation is to shed light on the different approaches followed by banks in modelling LGD and the different challenges and constraints this calculation faces.

Introducing Skewness into a Robust Model for Sequential Regression Multiple Imputation

Michael Johan von Maltitz
University of the Free State

A Bayesian paradigm is followed within sequential regression multiple imputation (SRMI), allowing for the use of various posterior predictive distributions for imputing missing values, based on the actual distributions of the incomplete variables. One concern within SRMI is misspecification of the imputation model. This concern would be somewhat alleviated if a robust model could be found that would provide adequate imputations regardless of the distribution of the underlying data.

This paper implements and assesses the robust skew t-distribution model in SRMI, as a replacement for the Normal, symmetric *t*, and other regression models. Additionally, alternative symmetric model adaptations used to incorporate skewness are compared with the actual skew *t* model.

Common Components To Construct Biplots For Longitudinal Data

Darryn Williams and SugnetLubbe
Department of Statistical Sciences, University of Cape Town

Statistics is well acquainted with the use of exploratory techniques to reveal detail about the structure of the data. Graphical displays of data are key in this process, chief amongst them the scatterplot. However, a scatterplot is restricted to use with data comprising at most three variables. For higher dimensional datacomprising $p > 3$ variables, the analogue to this display is the biplot. It is a plot that displays axes for all the variables that have been measured together with points to represent all the subjects in a low dimensional approximation of the p dimensional space. It is useful in that it affords the means to instantly assess correlations between variables as well as any groupings of the subjects that might be inherent in the dataset.

This display is well developed for data comprising *n* samples and *p* variables. Such datasets are referred to as two mode data where the objects and variables represent modes. This paper is focused on using biplots for data of a different nature. More specifically, three mode data is considered. Generally this type of data comprises a number of objects on which a number of measurements have been under different conditions. The modes are thus subjects, variables and conditions.

The dataset used in this paper is three mode by virtue of the fact that it is longitudinal in nature. It is taken from an investigation by Mansoor *et al.* (2009) on the immune response induced by Bacille Calmette-Guerin vaccine in infants exposed to the HIV. Cytokine expression in infants was measured at four different occasions.

The idea of common principal components (CPC) was introduced by Flury (1988). As opposed to doing separate principal component analyses on each of the data matrices comprising the three mode dataset, CPC seeks to find components common to all the data matrices. The method can be applied to data comprising independent groups as well as dependent groups. In the latter instance it is referred to as dependent common principal components (DCPC). The idea of commonality can also be extended to Canonical Variate Analysis (CVA). The main thrust of the paper is examining how the idea of commonality can be used in the construction of biplots for three mode data.

PROGRAMME

MONDAY: 05 NOVEMBER 2012		
08h00 - 09h00	**Registration**	
	Registration venue: Madibaz, South campus	
	Venue: Lab 5	Venue: Lab 4
09h00 - 11h00	WORKSHOP 1: Performance Measurement for Enterprises: An introduction for statisticians	WORKSHOP 2: Functional Data Analysis Using R
	Presenter: Prof Nicholas Fisher	Presenters: Prof Bhaswati Ganguli and Prof Agik Sen-Roy
	Session 1	Session 1
11h00 - 11h30	**Tea**	
	Venue: Madibaz	
11h30 - 13h00	WORKSHOP 1	WORKSHOP 2
	Session 2	Session 2
13h00 - 14h00	**Lunch**	
	Venue: Madibaz	
14h00 - 15h30	WORKSHOP 1	WORKSHOP 2
	Session 3	Session 3
15h30 - 16h00	**Tea**	
	Venue: Madibaz	
16h00 - 17h00	WORKSHOP 1	WORKSHOP 2
	Session 4	Session 4
17h30 for 18h00	**Meet and Greet**	
	Venue: Rendevouz, South Campus	

TUESDAY: 06 NOVEMBER 2012			
08h00 - 09h00	**Registration**		
	Registration venue: Madibaz, South campus		
	Venue: Lab 4	Venue: Madibaz, Umvelani	Venue: Lab 5
09h00 - 11h00	WORKSHOP 3: Nonparametric Methods with Instrumental Variables	WORKSHOP 4: SAS Workshop	
	Presenter: Prof Jeffrey Racine	Presenter: Mr Goran Dragosavac	
	Session 1	Session 1	
11h00 - 11h30	**Tea**		
	Venue: Madibaz		
11h30 - 13h00	WORKSHOP 3	WORKSHOP 4	
	Session 2	Session 2	
13h00 - 14h00	**Lunch**		
	Venue: Madibaz		
14h00 - 15h30	WORKSHOP 3	WORKSHOP 4	WORKSHOP 5: Education Committee
			Presenters: Prof Delia North and Ms Hannah Gerber
	Session 3	Session 3	Session 1
15h30 - 16h00	**Tea**		
	Venue: Madibaz		
16h00 - 17h00	WORKSHOP 3	WORKSHOP 4	WORKSHOP 5
	Session 4	Session 4	Session 2
17h15 - 18h00	SASA Exec meeting		
	Venue: Madibaz, Umvelani		
17h30 for 18h00	**Meet and Greet**		
	Venue: Rendezvous Café, South Campus		

WEDNESDAY: 07 NOVEMBER 2012			
07h30 – 12h00	**Registration**		
	Venue: North campus, Auditorium Foyer		
09h15 - 10h45	**Opening Ceremony**		
	Venue: North campus, Auditorium		
	09h15 - 09h20	MC: SASA President: Dr Gary Sharp	
	09h20 - 09h30	Welcoming: NMMU Dean of Science: Prof Andrew Leitch	
	09h30 - 10h00	Presidential address: SASA President: Dr Gary Sharp	
	10h00 - 10h30	Awards: SAS awards for best honours projects: Ms Hannah Gerber (hand over by Mr Murray de Villiers)	
		Awards: STATSSA awards for best postgraduate papers: Prof Delia North	
		Awards: ICCSSA award: Dr Pravesh Debba	
		Award: Sichel medal: Dr Roelof Coetzer (on behalf of Prof Paul Fatti).	
		Awards: Fellowship and Honorary Members: Dr Pravesh Debba	
	10h30 - 10h40	Platinum Sponsor Address: SAS, Mr Murray de Villiers	
	10h40 - 10h50	Survey analysis: Former President of SASA: Dr Herrie van Rooy	
10h55 - 11h25	**Tea**		
	Venue: North campus, Auditorium Foyer		
11h30 - 12h25	**Plenary Session**		
	Venue: North campus, Auditorium		
	11h30 - 12h25	Guest and Title	Prof Jeffrey Racine: Kernel Smoothing of Categorical Covariates: Theory and Application
		Chair	Dr Gary Sharp
12h45 - 14h00	**Lunch**		
	Venue: Madibaz		

WEDNESDAY: 07 NOVEMBER 2012					
14h00 - 16h00	**Parrallel Sessions**				
	Stream	General and Distribution theory	Analytics Competition	Biometry	Young statistician
	Chair	Dr Schalk Human	Dr Leonard Santana	Dr Kerry Leask	Mr Warren Brettenny
	Venue	5 0003	5 0005	5 0007	5 0002
14h00 - 14h15		Forecasting South African Aggregate Retail Sales Using Univariate Linear, Nonlinear, And Nonparametric Time Series Models	Bayesian Estimators Of The Location Parameter Of The Normal Distribution With Unknown Variance	An Overview Of Stepped Wedge Designs	The Discrete Pulse Transform For Images
		Goodness C. Aye,, Mehmet Balcilar and Rangan Gupta	Janet van Niekerk and Andriette Bekker	Kerry Leask	Inger Fabris-Rotelli
14h20 - 14h35		Comparing Predictions From Combined Nonparametric And Parametric Models For Estimating fCO_2 In The Southern Ocean	Youth Of South Africa: Comparing Those In Education, Training And Employment With Those Who Are Socially Excluded	Spatial Variation In Under-Five Mortality In South Africa	Detection And Down-Weighting Of Outliers In Non-Normal Clustered Data
		Wesley Pretorius and Sonali Das	Michael Medupi	Sulaiman Salau	Tinashe Chatora
14h40 - 14h55		Comparison Of Test Estimators For A Location Parameter From A Normal Model Under BLINEX Loss With Special Focus On Preliminary Test Estimation	Modelling Track Records Using Compound Distributions	Multi-step Approaches To Consumer Segmentation	Statistical Techniques To Calculate Loss Given Default
		Judy Coetsee , Prof. A Bekker and Dr. S.Millard	Erin Hanly, David Friskin and Gary Sharp	Marieta van der Rijst, Tormod Næs and Nina Muller	Jaco van Tonder
15h00 - 15h15		The Description Of The Extended Bimatrix Variate Beta Type II Distribution	Efficiency Of The MCMC With EM Algorithm As A Multiple Imputation Method	Designs for Field Trials with Unreplicated Treatments	On Systems With Gradual Repair
		Andriëtte Bekker, Kotie Roux, Karien Adamski and Schalk Human	Huibrecht Elizabeth Maartens	Linda M. Haines	Zani Ludick
15h20 - 15h35		Origin And Application Of A Noncentral Generalized Multivariate Beta Type II Distribution	Diagnostic Tests For The Distribution Of Random Effects	Optimal Designs For Two-Colour Microarray Using Linear Mixed Models	Common Components To Construct Biplots For Longitudinal Data
		Schalk Human, Karien Adamski, Andriette Bekker, Kotie Roux	Leonard Santana	Legesse K. Debusho and Dibaba B. Gemechu	Darryn Williams and Sugnet Lubbe
15h45 - 16h15	**Tea**				
	Venue: Madibaz				

WEDNESDAY: 07 NOVEMBER 2012					
16h15 - 17h15	**Parrallel Sessions**				
	Stream	Applied Statistics	Analytics Competition	Finance	Energy/Applied stats
	Chair	Ms Marien Graham	Dr Michael von Maltitz	Dr Franck Adekambi	Dr. Morné Rowan Sjölander
	Venue	5 0003	5 0005	5 0007	5 0002
16h15 - 16h30		A Comparison Of Three Phase I Control Charts	A Hierarchical Bayesian Model: Combining The Generalized Gamma And Generalized t-Distributions	Health Care Insurance Pricing	Forecastability, Stability and Cost of Forecast Error
		Margarethe Coelho	Stefan S Britz and Prof DJ de Waal	Franck Adekambi and Salha Mamane	Senzo Myeni
16h35 - 16h50		Nonparametric CUSUM And EWMA Control Charts For Monitoring Unknown Location Based On The Exceedance Statistic	Introducing Skewness into a Robust Model for Sequential Regression Multiple Imputation	Analyzing Financial Time Series With Varying Volatility And Extreme Clusters	Generating Guidance On Public Preferences For The Location Of Wind Turbine Farms In The Eastern Cap
		Mrs. Marien.A. Graham	Michael Johan von Maltitz	Frans F. Koning and Prof. D. de Waal	Jessica Hosking, Mario du Preez and Gary Sharp
16h55 - 17h10		Improved Probability Limits Design for Attribute Data	South African Pension Funds: A Review Of Construction Strategies And Empirical Evidence	Model Risk Management	Further Techniques In Estimating Temperature In An Area In Which Temperature Is No Longer Measured
		Sandile Shongwe	JD van Heerden	Carel van der Merwe	Dr. Morné Rowan Sjölander
16h15 - 17h15	**ICCSSA Board meeting**				
	Venue: 5 0001				
17h30 - 18h30	**SASA AGM**				
	Venue: 5 0007				
19h15 for 19h30	**Welcoming Function**				
	Venue: Humewood Golf Club				

THURSDAY: 08 NOVEMBER 2012					
07h30 - 09h00	**Young Statisticians Breakfast**				
	Venue: Blue Waters Café, Marine drive, Hobie Beach				
07h45 - 09h00	**Registration**				
	Venue: Foyer, building 5				
09h10 - 10h30	**Parrallel Sessions**				
	Stream	Biostats	Analytics Competition	Industrial statistics	General/Applied stats
	Chair	Dr Legesse Debusho	Mr Caston Sigauke	Dr Roelof Coetzer	Mr Mark Nasila
	Venue	5 0003	5 0005	5 0007	5 0002
09h10 - 09h25		Modelling Of Correlated Soil Animals Count Data	Measures Of Income Inequality	Accounting For Covariance Structure Changes Over Time In Multivariate Statistical Process Monitoring	Crime Mapping In Ekurhuleni
		Legesse Kassa Debusho and Gudeta W. Sileshi	Rarang Phillemon Dikgale, MR Makwela and Prof A Tessera	André G Mostert and Gerhard Koekemoer	Elsabe Smit
09h30 - 09h45		Non-Linear Regression Models For The Characterisation Of The Early Bactericidal Activity (EBA) Of Tuberculosis Drugs	Hawkes Processes And Their Financial Applications	Multivariate Statistical Process Monitoring By Combining PLS And PCA	One-way ANOVA with unequal variances
		Divan Burger and Robert Schall	Brendon M. Lapham and Iain L. MacDonald	Gerhard Koekemoer and Roelof LJ Coetzer	Henri Moolman
09h50 - 10h05		Latent Class Analysis in Substance use Research	Volatility Modelling Of Heteroskedastic Shares Using Generalised Pareto Distribution	Multivariate Economic Statistical Process Control	The Median Odds Ratio (MOR) for Interpreting Random Effects of a Multi-Level Logistic Regression Model
		Esmè Jordaan and Petal Petersen	Rhoda Makhwiting, Caston Sigauke, Maseka Lesaoana	Precious Mudavanhu and Pieta van Deventer	Trishanta Padayachee
10h10 - 10h25		SAS Sponsor	Volatility Modelling of the Producer Price Index in South Africa	Multivariate Statistical Process Evaluation By Procrustus Analysis For Complex Chemical Processes	Multinomial Logistic Regression to Predict a Merchant's Experience of Card Fraud Incidents
		Goran Dragosavac	Selaelo Lynette Ramare and Caston Sigauke	Roelof LJ Coetzer, Ruan F Rossouw and Niel le Roux	Mark Nasila and Gary Sharp
10h30 - 11h00	**Tea**				
	Venue: Madibaz				

THURSDAY: 08 NOVEMBER 2012					
11h00 - 11h55	**Plenary Sessions**				
	Venue: 5 0007				
	11h00 - 11h55	Guest and Title	Prof Nicholas Fisher: Statistics, Performance Measurement and Corporate Collapses. Opportunities for statisticians to have an impact.		
		Chair	Dr Roelof Coetzer		
12h00 - 13h00	**Parrallel Session**				
	Stream	Applied Statistics	Analytics Competition	Classification	Econometric application
	Chair	Prof Jacky Galpin	Prof Sugnet Lubbe	Mr Mark Dowdswell	Ms Debbie Lee
	Venue	5 0002	5 0005	5 0007	5 0003
12h00 - 12h15		Curable Shock Processes	Partial Least Squares Regression Biplots	High Dimensional (Penalised) Discriminant Analysis	Willingness-To-Pay For Improved Fish Stock Size And Levels In The Sundays River Estuary, Eastern Cape, South Africa: A Choice Experiment
		Maxim Finkelstein	O.V.F. Oyedele, S. Lubbe and N.J. Le Roux	Mark Dowdeswell, Tea Jashashvili, Kristian Carlson, Damiano Marchi, Robert Nshimirimana	Debbie Lee
12h20 - 12h35		Multivariate Spacings, Generalized Quantiles and Level Sets Quantiles	Biplots For Investigating Differential Living Conditions Between South African Racial Groups	Variable Selection And Binary Classification For Infrared Spectroscopy Data	It Rained But I Have No Rainfall Data!
		David Mason	Tsiresy Pierre Bernard and Sugnet Lubbe	Nelmarie Louw, Sarel Steel and Hélène Nieuwoudt	Nothabo Ndebele
12h40 - 12h55		Automatic Interaction Detection For Longitudinal Data	Aspects Of Multi-Label Classification	Biplots Constructed From Distance Matrices	A linear model for valuating preferences of freshwater inflows into forty selected estuaries along the South African coastline
		Jacky Galpin	Ivona Contardo-Berning and Prof. Sarel Steel	Niel Le Roux, Sugnet Lubbe and John Gower	Melnick Smith, Johan Hugo and Gary Sharp
13h00 - 14h00	**Lunch**				
	Venue: Madibaz				

THURSDAY: 08 NOVEMBER 2012					
14h00 - 16h00	**Parrallel Sessions**				
	Stream	Education	Analytics Competition	Bayesian	Biostats
	Chair	Prof Delia North	Dr James Allison	Dr Lizanne Raubenheimer	Prof Francesca Little
	Venue	5 0003	5 0005	5 0007	5 0002
14h00 - 14h15		Maths4Stats – To Be Involved Or Not To Be Involved – That Was The Question	Optimal Design For Two-Colour Cdna Microarray Experiment	Analysis Of Non-Food Household Expenditures Using Multivariate Structured Additive Regression Models	Meta-analysis of Covariate Effects Reported at Multiple Time Points
		Rénette J Blignaut	Dibaba B. Gemechu and Legesse K. Debusho	Lawrence N. Kazembe	Alfred Musekiwa, Prof. Samuel OM Manda and Prof. Henry Mwambi
14h20 - 14h35		Rewriting, Re-Registering But Not Always Revising: Patterns Of Behaviour In ODL Statistics Service Module Students	Comparison Of Some Methods For The Identification Of Common Eigenvectors	Bayesian Inference For The Piecewise Exponential Model Using Objective Priors	Frailty Modeling Of Recurrent Events
		Eeva Rapoo	Theo Pepler	A.J. van der Merwe and P.C.N. Groenewald	Jacob Majakwara
14h40 - 14h55		Investigating Success Rates Of First Level Statistics Students Before And After The Implementation Of An Online Homework System Called Aplia In 2012	Permutation Procedures for ANOVA, Regression and PCA	Comparison Of Objective Priors For The Censored Rayleigh Model	The Distribution of Data From Thorough QT Studies: - Normal or Log-Normal?
		Fransonet Reyneke and Lizelle Fletcher	Christine Storm and Dr Lizelle Fletcher	Johannes Theodorus Ferreira	Wallina Oosthuizen and Robert Schall
15h00 - 15h15		Reporting On Student Assessment of the Class Room Environment	A Review Of Non-Standard Applications Of SPC Charts	Adaptations Of KS Test With Applications	Smoothed Temporal Atlases Of Age-Gender And Cause Specific Mortality In South Africa
		Prof. Delia North	Mandla Diko	Sean van der Merwe and DJ de Waal	Samuel OM Manda
15h20 - 15h35		Challenges Of Large Class Teaching And Ideas To Facilitate A "Student-Centered" Learning Environment	Asymptotic Normality For The Nonparametric Estimator Of The Quintile Share Ratio	Comparing Two Poisson Means: A Bayesian Approach	Multivariate Spatial – Tempora Modeling of HIV and TB Co-Infections in South Africa
		Yoko Chhana	Tchilabalo Abozou Kpanzou and Tertius de Wet	Lizanne Raubenheimer and Abrie van der Merwe	Annah Managa

THURSDAY: 08 NOVEMBER 2012					
15h40 - 15h55		Students' Knowledge And Perception Of Quality Control Measures In Educational Services In Yaba College Of Technology, Lagos, Nigeria	Hidden Markov Model Extensions For Animal Movement Models	Using The Markov Chain Monte Carlo Method To Make Inferences On Items Of Data Contaminated By Missing Values	Models For The Analysis Of Competing Risks
		Adeyemi Davidson Aromolaran	Victoria Goodall, Paul Fatti and Norman Owen-Smith	Innocent Karangwa and Prof D Kotze	Francesca Little
16h00 - 16h30	**Tea**				
	Venue: Madibaz				
16h30 - 17h30	**Parrallel Sessions**				
	Stream	Applied Statistics	Analytics Competition	Process control	Cointegration
	Chair	Dr Tom Berning	Ms Inger Fabris-Rotelli	Ms Michelle Botes	Dr Vince Micali
	Venue	5 0002	5 0005	5 0007	5 0003
16h30 - 16h45		Multivariate Nonparametrics In Practice	An Overview Of Noise In Signal Analysis	Statistical Quality Control With Autocorrelated Data	Application of Granger Causality to Energy Market Models in South Africa
		Prof Chris F Smit	PK Smith and IN Fabris-Rotelli	Michelle Botes	Vanessa Ndlovu
16h50 - 17h05		Partial Least Squares (PLS) Variable Selection Using A Hybrid Particle Swarm Optimization Algorithm	An Overview Of Image Segmentation Techniques	A Comparison Of The Minimum And The Precedence Charts For Subgroup Data	Examining The Link Between Unemployment And Inflation Using Johansen's Co-Integration Approach And Vector Error Correction Modelling
		Martin Philip Kidd and Martin Kidd	J-F Greeff and IN Fabris-Rotelli	Jean-Claude Malela-Majika	Sagaren Pillay
17h10 - 17h25		Quantification of Estimation Instability and its Application to Threshold Selection in Extremes	An Investigation And Historical Overview Of The G/M And M/G Queueing Processes	Cumulative Sum (CUSUM) Control Chart: Refinements And Enhancements	Examining Long-Run Relationships Between The BRICS Stock Market Indices To Identify Opportunities For Implementation Of Statistical Arbitrage Strategie
		Tom Berning	C Kraamwinkel and IN Fabris-Rotelli	Frederik Coenrad Delport	Brian Meki and D Kotze
18h30 for 19h00	**Gala Dinner**				
	Venue: The Willows				

FRIDAY: 09 NOVEMBER 2012					
08h50 - 09h30	**Registration**				
	Venue: Foyer, building 5				
09h10 - 10h30	**Parrallel Sessions**				
	Stream	Energy	Official Statistics/Classification	Applied Statistics	Applied Statistics
	Chair	Dr Delson Chikobvu	Dr Pravesh Debba	Prof Maseka Lesoanna	Dr Lizelle Fletcher
	Venue	5 0003	5 0005	5 0007	5 0002
09h10 - 09h25		Modelling Daily Increases In Peak Electricity Demand Using The Generalized Pareto Distribution	Sampling Design And Analysis In The Cardiovascular Risk Household Survey	Numerical Maximization Of Likelihoods, E.G. As An Alternative To EM	Panel Data Regression
		Caston Sigauke, Delson Chikobvu, Andréhette Verster	Nomonde Gwebushe , Carl Lombard, Nasheeta Peer, Naomi Levitt, Krisela Steyn, Estelle Lambert	Iain L. MacDonald	Dr Lizelle Fletcher, Ms S Surovitskikh, Prof B Lubbe
09h30 - 09h45		Tail Quantile Estimation Of Heteroskedastic Intraday Increases In Peak Electricity Demand	Herbaceous Biomass Prediction From Environmental And Remote Sensing Indicators	Accuracy And Fairness Of Rain Rules For Interrupted One-Day Cricket Matches	The Effect of Differential Skewness on the Factor Structure of Equal Interval Scale Data
		Caston Sigauke, Andréhette Verster and Delson Chikobvu	Nontembeko Dudeni-Tlhone, Abel Ramoelo. Pravesh Debba, Moses Azong Cho, Renaud Mathieu	Robert Schall and Dianne Weatherall	Jaclyn Smith and Dr Lizelle Fletcher
09h50 - 10h05		Predicting Zimbabwe's Annual Rainfall Using Darwin Sea Level Pressure Index	Combining Binary Classifiers To Improve Tree Species Discrimination At Leaf Level	Using P-values for Multiple Comparisons	Bayesian Hierarchical Spatiotemporal Models In Epidemiology: A Case Study Of Tuberculosis In Kenya
		Retius Chifurira and Delson Chikobvu	Xolani Dastile, Gunther Jäger, Pravesh Debba, and Moses Cho	Francois Steffens	Thomas Achia
10h10 - 10h25		Exact Confidence Intervals For The P Quantile Based On Order Statistics Of A Two-Parameter Weibull Distribution	The MDS-GUI: A Graphical User Interface For Comprehensive Multidimensional Scaling Applications	Comparative Analysis Of Price Indices Of Electrical Appliances In South Africa	Aspects Of The V-Soft Minimal Hypersphere As An Outlier Detector For Multivariate Data
		Peter Iiyambo	Andrew Timm and Sugnet Lubbe	Happy Maluleke, Maseka Lesaoana, Rahab Makwela	Morné M.C. Lamont

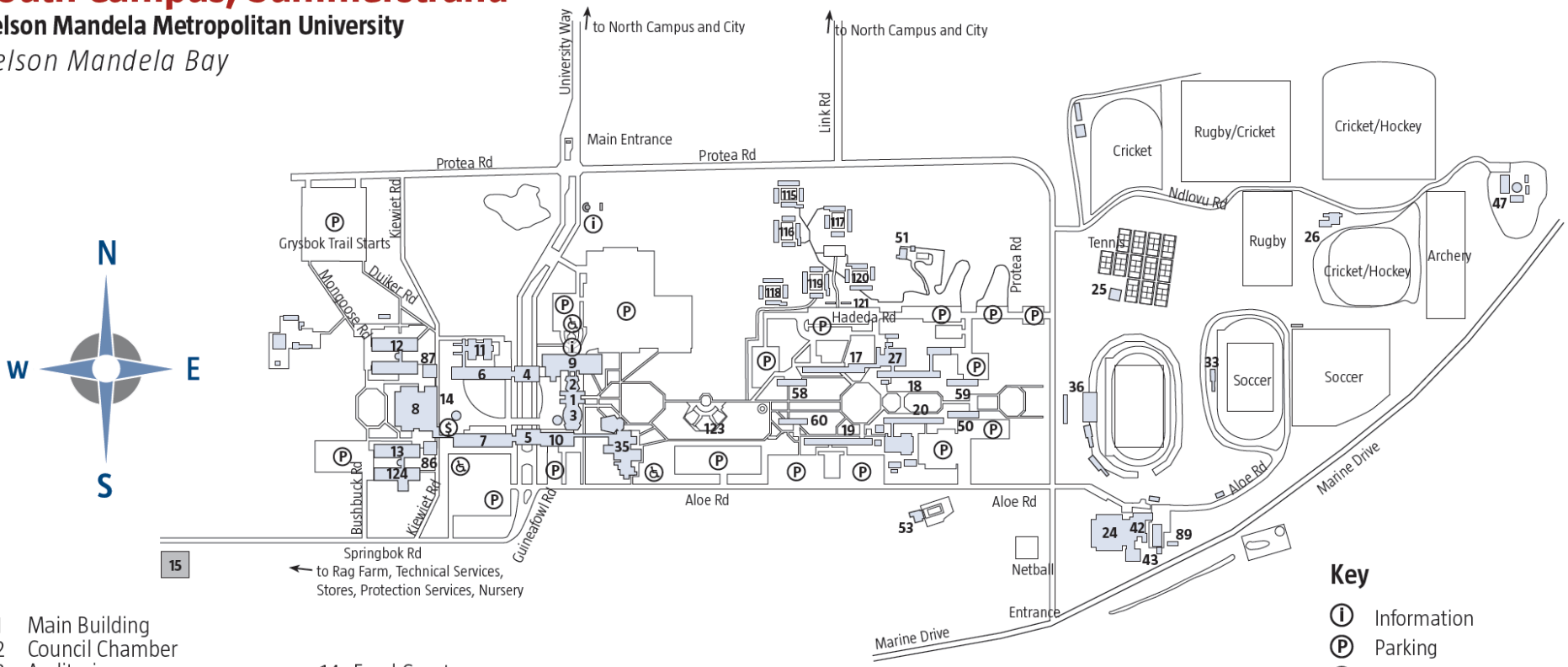
FRIDAY: 09 NOVEMBER 2012			
10h30 - 10h55	**Tea**		
	Venue: Madibaz		
11h00 - 11h55	**Plenary Session**		
	Venue: 5 0007		
	11h00 - 11h55	Guest and Title	Prof Tertius de Wet - Campanometry
		Chair	Prof Delia North: Chair of SASA Education Committee
12h00 - 12h30	**Closing: SASA 2012**		
	Dr Roelof Coetzer		
	Venue: 5 0007		

MAPS

South Campus, Summerstrand

Nelson Mandela Metropolitan University

Nelson Mandela Bay



1 Main Building
2 Council Chamber
3 Auditorium
4 Old Mutual Lecture Halls
5 Sanlam Lecture Halls
6 Education, Writing Centre & ABSA Computer lab
7 M & P Building
8 Library & School of Architecture
9 Embizweni
10 Music
11 Human Movement Sciences & Biokinetics Centre
12 Biological Sciences
13 Physics & Chemistry
14 Food Court
15 Technical Services
17 Unitas Main Block
18 Veritas Main Block
19 Xanadu Main Block
20 Melodi Main Block
24 Indoor Sport Centre & Sport Offices
25 Tennis Clubhouse
26 Cricket Clubhouse
27 Study Centre (Veritas)
33 Soccer Clubhouse
35 Building 35 (Universet Lecture Halls)
36 Stadium & Clubhouse
47 Alumni Centre (Campus Boma)
50 Melodi Annex
51 Unitas/Veritas Clubhouse & Pool
53 Xanadu/Melodi Clubhouse & Pool
58 Unitas Annex
59 Veritas Annex
60 Xanadu Annex
89 Underwater Clubhouse
86 Goldfields South
87 Goldfields North (International Office)
115-120 Renaissance Postgrad Student Village
121 Housing Administration
123 Building 123
124 Higher Resolution Transmission Electron Microscope Centre (HRTEM)

Key

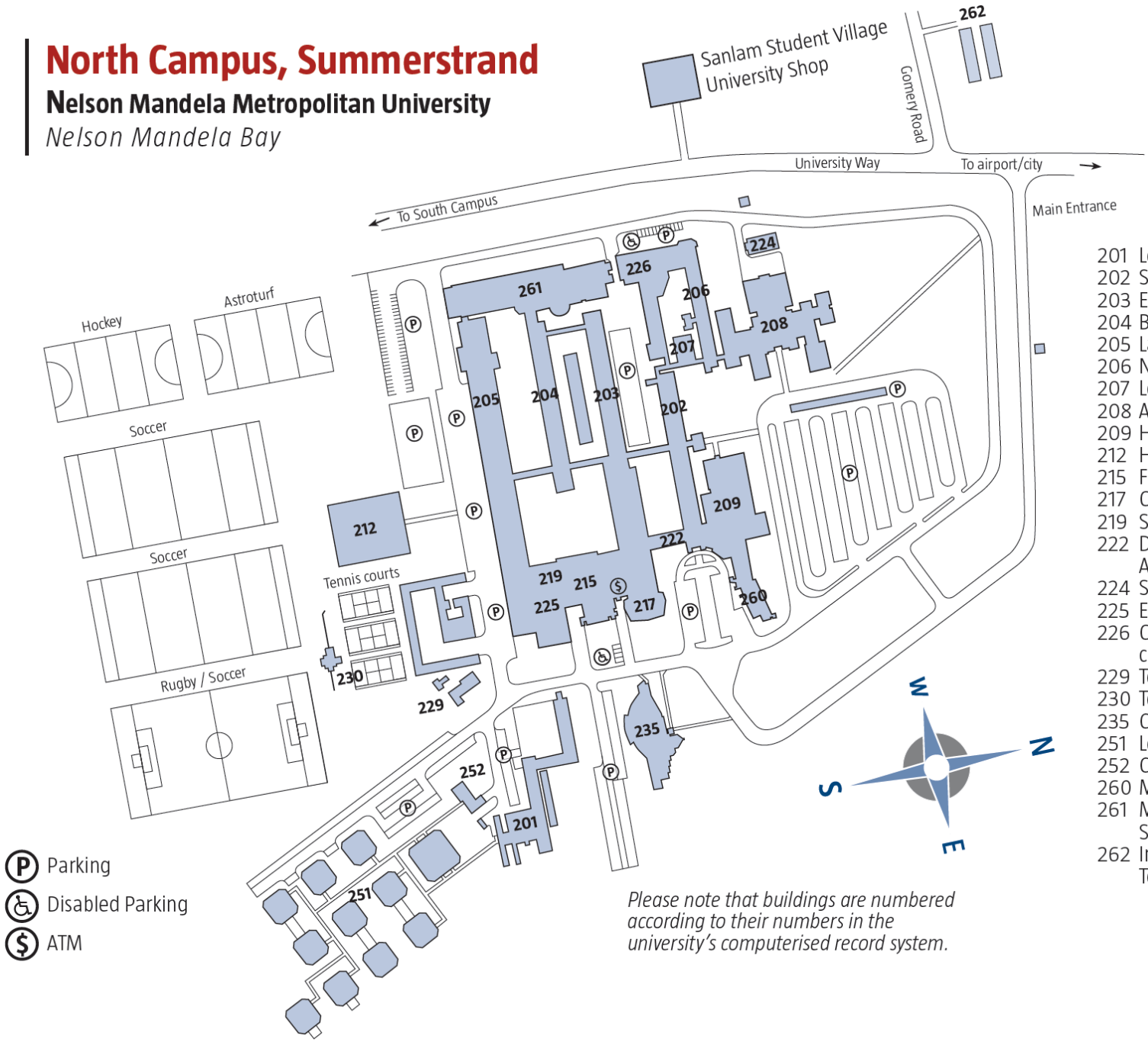
- Information
- Parking
- Disabled Parking
- ATM

Please note that buildings are numbered according to their numbers in the university's computerised record system.

North Campus, Summerstrand

Nelson Mandela Metropolitan University

Nelson Mandela Bay



- (P) Parking
- Disabled Parking
- ($) ATM

201 Lebombo Residence
202 Science & Health Sciences
203 Engineering
204 Built Environment & Engineering
205 Laboratories
206 Nursing
207 Lecture Halls (N1 & N2)
208 Art
209 Human Resources & Contact Centre
212 Heinz Betz Centre
215 Food Court (cafeteria)
217 Goldfields Auditorium
219 Student Services Centre
222 D Block (main entrance, cashiers, ABSA Alumni Centre, Student Counselling)
224 Stained Glass Studio
225 EBEIT Lab
226 Goldfields Computer Centre (writing centre, Offices etc.)
229 Technical Services
230 Tennis Clubhouse
235 Conference Centre
251 Letaba Residence
252 Campus Health Services
260 Main Administration Building
261 MTLC Building (eNtsa, library, CTLM, Senate Hall, lecture halls)
262 InnoVenton / Downstream Chemical Technology Station

Please note that buildings are numbered according to their numbers in the university's computerised record system.