ABSTRACTS: 20th Australian Statistical Conference 2010 Fremantle, WESTERN AUSTRALIA

including **OZCOTS 2010** - The 7th Australian Conference on Teaching Statistics

Abstracts of the **ASC presentations** are included by alphabetical / presenting author.

- Abstracts for **Invited Speakers** commence on page 31.
- Abstracts for **Concurrent Sessions** commence on page 45.
- Abstracts for **Posters** commence on page 259.

Abstracts of the **OZCOTS presentations** are included in program order from page 287.

An **Author Index** commences on page 313.

ASC 2010 SCIENTIFIC PROGRAM COMMITTEE

Ray Chambers (University of Wollongong)
Brenton Clarke (Murdoch University) *Chair*
Ross Darnell (CSIRO)
John Henstridge (Data Analysis Australia) *Deputy Chair*
Charles Pearce (Adelaide University)
Katia Stefanova (Department of Agriculture and Food, WA)
Paul Sutcliffe (Australian Bureau of Statistics)
Alan Welsh (Australian National University)
Frank Yu (Australian Bureau of Statistics)

The ASC Abstracts were reviewed and edited by Brenton R Clarke with assistance from the following people: Russell John, Anna Munday, Katia Stefanova, Jane Speijers Ross Taplin, Berwin Turlach, Nihal Yatawara and Frank Yu.

CRISP REVISITED: EFFECTIVENESS AND EFFICIENCY IN AUSTRALIAN OFFICIAL STATISTICS

Geoff Allen, AM

Allen Consulting Group
Level 9, 60 Collins Street, Melbourne, Victoria
mdowel@allenconsult.com.au

The 1974 (Crisp) Committee on Integration of Data Systems in Australian government concluded that there was serious sub-optimisation of the statistical efforts across departments and levels of government, inefficient use of the scarce statistical expertise available to the official statistics community, fragmentation, duplication and incompatibility of decentralised uncoordinated statistical collections and unnecessary response burdens on providers of data. It also noted an inadequate use across government of administrative data due to a lack of standardisation across agencies and a lack of data integration.

This was seen to limit the value of available data for policy making, as well as being wasteful of financial and human resources. Significant administrative and policy changes were introduced to address these concerns.

A current assessment is offered suggesting that many of these deficiencies continue to exist or have re-emerged with a consequent sub-optimisation of evidence based decision making and inefficient use of scarce expertise and financial resources.

While technical advances are facilitating improvements, the complexities of modern government are making co-ordination and data integration more difficult. However new demand is being felt for data collection and analysis in areas such as social inclusion, environment including impacts of climate change policies, and infrastructure, which will keep pressure on scarce government resources and threaten to increase ad hoc and suboptimal statistical initiatives in agencies.

Some directions for the way ahead are made.

***Geoff Allen, AM**, was senior advisor to the Federal Treasurer and Leader of the Opposition, business school academic and co-founder and foundation CEO of the Business Council of Australia. He is founder and is currently Director of the Allen Consulting Group. He has been Chairman of a number of State and Commonwealth Government advisory councils including the Australian Government's Trade Policy Advisory Council and director of a number of public companies. He is currently Chairman of the Australian Statistics Advisory Council, National Chairman of the Committee for Economic Development of Australia (CEDA) and Chairman of the Australasian Centre for Corporate Public Affairs.*

NEW METHODOLOGY FOR SPATIAL POINT PATTERNS WITH GEOLOGICAL, ENVIRONMENTAL AND ASTRONOMICAL APPLICATIONS

Adrian Baddeley

CSIRO, CMIS, Leeuwin Centre
65 Brockway Rd, Floreat, WA 6014
Adrian.Baddeley@csiro.au

The statistical analysis of spatial patterns of points has long been an important problem in astronomy, ecology and geology. This talk describes new statistical methodology for spatial point process models, and recent applications to the analysis of large point pattern datasets in these three areas. This work is a collaboration between CSIRO, The University of Western Australia, and Aalborg University.

We have developed analogues, for spatial point process models, of the classical techniques for model validation in (generalized) linear models such as residuals, leverage, influence measures, partial residual plots, added variable and constructed variable plots. We report an application to ecological surveillance of in Western Australian native woodland, on a scale involving millions of trees.

Classical methods for spatial point pattern analysis are based on empirical functional summary statistics such as the K-function and the two-point correlation function. We have developed a general approach, based on the score test, which extends these techniques to non-stationary processes, using a counterpart of the one-dimensional point process compensator.

In Geographical Information Systems (GIS), spatial point pattern data are often analysed by discretising space and applying logistic regression. It is not widely understood that this is equivalent to assuming a loglinear Poisson point process model. Benefits of this insight include better parameter estimation and more informative prediction. We demonstrate an application to prospective geology in Western Australia.

Adrian Baddeley *is one of Australia's leading statisticians. He is a winner of the Pitman Medal, Hannan Medal, and Australian Mathematical Society Medal, and is a Fellow of the Australian Academy of Science. He has held positions at Trinity College Cambridge, The University of Bath, CSIRO, the Centre for Mathematics and Computer Science in Amsterdam, and the University of Leiden in The Netherlands. Until recently he was Professor of Statistics at the University of Western Australia. He is now a Science Fellow with the CSIRO Division of Mathematics, Informatics and Statistics (CMIS) based at Floreat, Western Australia.*

STATISTICAL INFERENCE BASED ON FRÉCHET DIFFERENTIABILITY AND APPLICATIONS

Tadeusz Bednarski

Economics Institute, University of Wroclaw, Poland
ul. Uniwersytecka 22/26, 50-145 Wroclaw, Poland
t.bednarski@prawo.uni.wroc.pl

A key step in building the inference procedures for statistical models is in providing distributional properties of statistics. The task may become complex when estimates are given implicitly as it is so frequently with robust procedures. The method of Fréchet differentiability is then a reliable and relatively simple tool in deriving approximate distributions of estimators. An estimator depending on the empirical distribution function which is Fréchet differentiable is also robust in a qualitative sense – its influence function is bounded and smooth. Even though the differentiability does not provide efficiency, in conjunction with a general idea of smooth trimming of the likelihood ratio, it may lead to very reasonable robust procedures in nonstandard and complex inferential situations.
The aim of the lecture is to demonstrate that indeed the above general idea can be implemented usefully in practice. A differentiable modification of the partial likelihood estimator for the Cox model and a robust modification of the maximum likelihood estimator for the generalized Poisson model will be discussed and supplemented by real data examples (Bednarski (1993, 1996, 2004)). New results concerning the unit root testing in the time series will also be given to demonstrate potential use of the differentiability notion in statistical inference for stationary processes.

References

Bednarski, T. (1993). Robust estimation in Cox's regression model. Scandinavian Journal of Statistics. Vol. 20, pp 213-225.
Bednarski, T., Zontek, S. (1996). Robust estimation of parameters in a mixed unbalanced model. Annals of Statistics. Vol. 24, pp 1493-1510.
Bednarski, T. (2004). Robust estimation in the generalized Poisson model. Statistics, Vol. 38, pp 149-159

Tadeusz Bednarski *currently works at Economics Department of the University of Wroclaw as Head of Statistics Section. His main research interests are in asymptotic robust methodology, in particular in applications of the Fréchet differentiability to statistical inference in Poisson, Cox and time series models. From the practical standpoint his interests concern longitudinal studies of clinical data, insurance statistics, unemployment statistical studies, and econometric modelling in time series.*

PLANNING FOR HEALTH

Dr Jim Codde

A/Executive Director, Service Planning and Development, South Metropolitan Area Health Service
Western Australian Department of Health
Jim.Codde@health.wa.gov.au

The population of Western Australia has grown from 1.90 million in 2001 to 2.27 million in December 2009, with an annual rate of growth increasing from 1.4% to 3.1% over the same period, making WA's population the fastest growing in Australia.

But what does all this mean for the planning of public health services? Does the fact that overseas migration accounts for almost two-thirds of this growth impact on the health needs of our population? Despite the minerals led economic boom in the northwest of WA, how do we ensure the health infrastructure plans account for the fact that most of the population growth occurs elsewhere in the state? With the number of births in WA showing the largest proportional increase (9.2%) in births of all states and with total fertility rates in 2008 (2.12 babies per woman) at its highest rate since 1976, how are the future health service needs of the population estimated to meet demand, be located close to those at need and be delivered by an appropriately skilled workforce?

This presentation will outline how determining the current and future health needs of the community builds on a range of major statistical initiatives that include:

- National Health Dictionary
- National Coding Standards
- Health and Wellbeing Surveys
- Data linkage
- Population projections
- Clinical reforms in service delivery, and
- Data modelling.

The output of statistical demand modelling is used in the planning and building of billions of dollars worth of hospital infrastructure, the establishment of new and expanded tertiary education curricula to deliver the necessary clinical workforce, and major health reforms including activity-based funding. It is clear that these statistics have a major impact on health service planning and policy.

This presentation will attempt to touch on these aspects of data analysis and open the door to a lively discussion on how they impact on the way health services are delivered in Australia in an ever-changing environment.

Jim Codde *is currently Acting Executive Director of Service Planning and Development for the South Metropolitan Area Health Service in Western Australia. In this capacity he works with others in his Directorate and elsewhere within the WA Department of Health on the development of two new tertiary hospitals, the upsizing of smaller metropolitan secondary hospitals into large general hospitals and planning for one of the largest concurrent movement of patients and services from multiple hospitals when these facilities are commissioned in 2014. Jim is an Adjunct Professor with Curtin University and maintains strong links with the public health schools at other academic institutions. He has sat on a number of national committees and steering groups for the Australian Burden of Disease studies and been a member of the ABS's WA Statistical Policy Committee for several years.*

COMBINING OUTPUTS FROM AN ENSEMBLE OF REGIONAL CLIMATE MODELS

Noel Cressie

Program in Spatial Statistics and Environmental Statistics,
Department of Statistics, The Ohio State University
1958 Neil Avenue, Columbus, OH 43210-1247, USA
ncressie@stat.osu.edu

Regional climate models (RCMs) are important tools to study local climate behavior. The North American Regional Climate Change Assessment Program (NARCCAP) is an international program designed to provide high-resolution, climate-model output for the United States, Canada, and northern Mexico. RCMs use physical relationships, in subtly different ways, to downscale information from a coarse-resolution global climate model. Although all RCMs generally capture the large-scale spatial variation from coast to coast, from south to north, and from low to high elevations similarly, their outputs can differ substantially in some regions. In this talk, I analyze the 20-year-average Boreal winter temperatures from each of six RCMs that were run in Phase I of NARCCAP (where all RCMs are driven by the boundary conditions provided by the NCEP/DOE Reanalysis II data).

A Bayesian hierarchical model (BHM) is built, which includes a spatial meta-analysis model that allows a consensus regional climate to be defined from the ensemble of RCMs. In the BHM, each RCM's "vote" for the consensus climate is "counted" differently. An MCMC implementation enables posterior inference on the "unknowns," including the large-scale fixed effects and the small-scale random effects in each RCM and in the consensus climate.

Posterior inference on the resulting spatial random fields allows a visual comparison of the RCMs and the consensus climate. Because the BHM uses a fixed-rank model, Bayesian computation on the large datasets from the six RCMs is feasible. Additionally, the model has a spatial covariance structure that can capture the nonstationarities expected over a region with very heterogeneous physical geography. This talk is based on research with Emily Kang (SAMSI) and Steve Sain (NCAR).

Noel Cressie *received the Bachelor of Science degree with first class honours in Mathematics from the University of Western Australia. He received the MA and PhD in Statistics from Princeton University. Dr Cressie is Professor of Statistics, Distinguished Professor of Mathematical and Physical Sciences, and Director of the Program in Spatial Statistics and Environmental Statistics at The Ohio State University. His research interests are in the statistical modelling and analysis of spatial and spatio-temporal data. This has led to the development of Bayesian and empirical Bayesian methodology in complex, non-linear systems in the earth sciences, such as spatial analysis of mineral properties of soil, long-lead forecasting of the El Nino phenomenon, remote sensing of global environmental processes, and estimating ice-stream dynamics. He is the author of two books, including "Statistics for Spatial Data, rev. edn.", published by John Wiley and Sons. Dr Cressie is a Fellow of the American Statistical Association and the Institute of Mathematical Statistics, and he is an Elected Member of the International Statistical Institute. He was the 2009 Fisher Lecturer awarded by the Committee of Presidents of Statistical Societies.*

DEVELOPMENT POLICY IN A WORLD OF IMPERFECT INFORMATION

Nicky Cusworth

Department of State Development
Western Australia

In an ideal world, government decision making would be informed by a suite of information that is accurate, unbiased, timely and relevant. In the real world, the data available to inform our decisions is sometimes none of those things.

This presentation will look at the limitations of data available to inform state development policies, the problems those limitations present, and some possible ways around them. The problems discussed will include:

- Data limitations – when the information we need is not available or of poor quality
- Information asymmetry – when we negotiate with people who know more than we do
- Source shopping – picking the information that suits the story we want to tell
- Filling the vacuum – using bad data when good data isn't available

Nicky Cusworth *is Deputy Director General for Strategic Policy in the WA Department for State Development. She has extensive experience in economic analysis and public policy in the private and public sectors.*

ESTIMATING *N* PARAMETERS FROM A SAMPLE OF SIZE ONE: AN INTRODUCTION TO RANDOM NETWORKS

Persi Diaconis

Department of Statistics, Stanford University
390 Serra Mall, Stanford, CA 94305-4065, USA
diaconis@math.stanford.edu

Network data has become readily available and working scientists are building a slew of models to help make sense of it. As one thread through this, I will discuss so-called "power law" or "scale free" graphs. A natural model here has the degree sequence as sufficient statistics. I show that, with high probability, the MLE gets uniformly close to all n parameters, based on a single observed graph. I will try to explain the relevant graph theory---particularly the emerging theory of "graph limits" in "statistical English." This is joint work with Joe Blitzstein, Sourav Chatterjee, Susan Holmes, Svante Jansson, and Alan Sly.

Persi Diaconis *is the Mary V. Sunseri Professor of Statistics and Mathematics at Stanford University. Diaconis was born into a family of professional musicians. At 14 he quit his violin lessons at Julliard after 9 years of study, and went on tour with Dai Vernon, "the greatest magician in the US." Diaconis did well doing magic, inventing tricks, giving lessons and living a "very colorful" life for 8 years until he was recommended a probability book by Feller as the best and most interesting on the subject. Diaconis bought it and then found that he couldn't read it. So he enrolled in N.Y. City College at night, graduated two years later with a degree in mathematics and was accepted into the statistics program at Harvard. By 1974 he had earned a Ph.D. and joined the faculty of the Statistics Department at Stanford.*

IMPORTANCE SAMPLING: AN ALTERNATIVE VIEW OF ENSEMBLE LEARNING

Jerome H. Friedman

Department of Statistics, Stanford University
390 Serra Mall, Stanford, CA 94305-4065, USA,
jhf@stat.stanford.edu

Learning a function of many arguments is viewed from the perspective of high-dimensional numerical integration. It is shown that many of the popular ensemble learning methods can be cast in this framework. In particular, bagging, boosting, and Bayesian model averaging are seen to correspond to Monte Carlo integration methods each based on different importance sampling strategies.

This interpretation explains some of their properties and suggests modifications to them that can improve their accuracy and especially their computational performance.

This is joint work with Bogdan Popescu.

Jerry Friedman *is a pioneer in the theory and practice of computational statistics and data mining. He has been a Professor of Statistics at Stanford University for more than 20 years and has published on a wide range of data-mining topics including nearest neighbour classification, logistical regressions, and high dimensional data analysis. His primary research interest is in the area of machine learning. He has written many expository articles and books and given an extraordinary number of invited talks relating data mining and machine learning to statistical foundations, and developed and implemented new methodologies including CART, MARS, PRIM, PPR, MART, and Gradient Boosting He has won numerous awards for his many contributions to mathematical physics, and statistics. Jerry is a Fellow of the American Academy of Arts and Sciences, and a recipient of the prestigious Parzen Prize which is awarded to North American statisticians who have made outstanding and influential contributions to the development of applicable and innovative statistical methods. He is also a recipient of the ACM Data Mining Lifetime Innovation Award. A number of papers written by Jerry have been recognised by Technometrics and the Journal of the American Statistical Association as Paper of the Year. His attendance at ASC2010 has been sponsored by CSIRO.*

TRUSTWORTHY STATISTICS – A SHARED RESPONSIBILITY?

Denise Lievesley

School of Social Science and Public Policy
King's College London, London WC2R 2LS
denise.lievesley@kcl.ac.uk

Statistics which are both trustworthy and trusted are a pre-requisite for a healthy society. Recent years have seen the introduction of performance monitoring in many countries and statistical systems are increasingly called upon to support this process. Data are needed to establish 'what works' among policy initiatives; to identify those aspects of public services which perform well or poorly; and, equally important, to hold public servants and elected representatives to account. We have also witnessed the rise of such performance monitoring at international levels, where there can be big incentives for governments to paint a particular picture with the numbers.

Thus governments use statistics to monitor public services, whilst at the same time themselves being monitored by performance indicators. It is precisely because of this dual role that performance monitoring must be conducted with integrity and shielded from undue political influence.

The production of policy-relevant statistics which are politically independent is becoming ever more challenging. Building systems which meet these criteria is much too important to be left to official statisticians. Likewise a statistically-literate citizenship is essential for the building of an informed society - and this is too important to be left to statistical educators.

The statistical community incorporates practitioners in a wide range of application fields as well as educators and those developing the theoretical underpinnings of the discipline. With its unique breadth of expertise, this community can, working together, inform the public debate on the political and educational challenges of official statistics. Denise will highlight the role of statistical societies in creating this unity from diversity.

Denise Lievesley *is a Professor of Social Statistics and Head of the School of Social Science and Public Policy at King's College London. Professor Lievesley is one of the country's leading social statisticians, who has campaigned for evidence to be used as the basis for the development of sound public policies within the UK and more widely. Having enjoyed a distinguished career, which has included the posts of founding Chief Executive of the English Information Centre for Health and Social Care; Director of Statistics at UNESCO –where she established its new Institute for Statistics –and Director of the UK Data Archive (and simultaneously Professor of Research Methods in the Mathematics Department, University of Essex), most recently Professor Denise Lievesley was a special advisor at the African Centre for Statistics of the UN and was based in Addis Ababa. Professor Lievesley's various roles have led her to work with ministers, ambassadors, senior civil servants and officials of international agencies, for which she has established a reputation for upholding the principles of professional integrity, policy relevance and methodological transparency. Throughout her working life, Professor Lievesley has been committed to protecting the integrity of official statistics and to ensure that they remain free from political influence.*

BETTER USE OF INFORMATION IN GOVERNMENT

David Smith

Department of the Premier and Cabinet
Western Australia

Government is big business in its own right and has an even bigger impact through services delivered and laws and regulations affecting all our lives. They are also big collectors of data. Expectations are growing for what governments will deliver, how they will be accountable and how different levels and parts of government will work together seamlessly. The raft of recent COAG agreements between all Australian Governments is good example of that. Decisions made by governments must be well informed. Information is needed to identify what the problem is; what will fix it; whether a solution works; and who is accountable. Information can come from many sources, but government advisers need to know how to use it and have systems to manage it. This is a challenge that is being met within government but more could be done.

David Smith *is Deputy Director General at the Department of the Premier and Cabinet in the WA Government. He is responsible for Coordination, Cabinet and Policy Division. He has been in this position since August 2008. Prior to this he was a member of the corporate executive of the Department of Treasury and Finance, with responsibility for economic policy. Prior to this he had over 20 years experience in the Commonwealth public service, including in the Prime Minister's department and overseas postings with the Department of Foreign Affairs and Trade. He has also worked with a private economic consultancy in London.*

STATISTICAL INFERENCE WITH WHOLE-GENOME TRANSCRIPTIONAL DATA

Gordon K Smyth

Walter and Eliza Hall Institute of Medical Research
1G Royal Parade, Parkville, Victoria 3052
smyth@wehi.edu.au

Modern technologies such as microarrays and Next-Generation deep-sequencing platforms are able to measure the transcriptional level of every gene in the genome given a sample of RNA from a cell or tissue. This gives a snapshot of the activity level of every gene in a particular cell type at a particular time. These activity snapshots can then be related to various predictors such as disease status or genotype. In this way, we can study which genes are associated with different conditions, and hence learn much about gene function and inter-gene networks. This is an exciting playground for a statistician, not just because of the access to cutting-edge science, but also because of the opportunity to apply and develop a wealth of interesting statistical ideas. Genomic data is hugely high-dimensional with tiny sample sizes, so the familiar rules of univariate inference are often radically changed. This talk will describe some of our work using a variety of multivariate and empirical Bayes techniques to analyze transcriptional profiles, focusing particularly on RNA sequencing data from the latest technologies.

Gordon Smyth *is an NHMRC Senior Research Fellow and head of a research group working on statistical functional genomics in the Bioinformatics Division of the Walter and Eliza Hall Institute of Medical Research. He is well known internationally for his work on microarray gene expression data analysis.*

WHAT I SEE IS NOT QUITE THE WAY IT REALLY IS

Chris J Wild

Department of Statistics, University of Auckland
38 Princes St, Auckland, New Zealand
c.wild@auckland.ac.nz

In this talk we will gaze through the ripple glass of a bathroom window and wander Alice-like down garden paths through a wonderland where what we see is never quite the way it really is. The paths our odyssey leads us along are conceptual pathways that start with conceptualisations of statistical inference that are intended to be accessible to, and operable by, students mid-way through high school and lead us, via a series of connected trails, all the way to plot annotations that better reveal the stories being told by factor variables in generalised linear models. Along the way, both motivating and suggesting ways forward for all of this, we will meet novel visualisations of sampling variation, re-sampling variation and randomisation variation. The talk will draw on a paper with Maxine Pfannkuch, Matt Regan and Nicholas Horton entitled, “Towards more accessible conceptions of statistical inference” to be read to the Royal Statistical Society late in 2010 and on other work on making inference more accessible, particularly via visualisations, with these and other collaborators.

Chris Wild *- Professor of Statistics at the University of Auckland, New Zealand and recognised by Fellowships of the American Statistical Association and the Royal Society of New Zealand, Chris Wild is a member of a rare crossover species. He publishes extensively in statistical methodology, particularly on response-selective and missing data problems, but also works substantively in statistics education. He co-wrote the Wiley books Nonlinear Regression (1989) and Chance Encounters (2000) with George Seber. His best known statistics education paper is Statistical Thinking in Empirical Enquiry with Maxine Pfannkuch (1999, International Statistical Review). Chris' interests in statistics education include curricular revolution at school levels, growing university statistics programmes, and improving the penetration, quality and practical impact of statistics education at all levels. Chris has been a Council member of the International Statistical Institute, President of the International Association for Statistics Education and an Associate Editor of the International Statistical Review, Biometrics, the Statistics Education Research Journal, and ANZJS. He was Head of Auckland's Department of Statistics 2003-2007 and co-led the University of Auckland's first-year statistics teaching team to a national teaching award in 2003. His keynote addresses include the Royal Statistical Society, the Statistical Society of Canada, and ICOTS.*

USING HIERARCHICAL MODELS TO ATTRIBUTE SOURCES OF VARIATION IN CONSUMER ASSESSMENTS OF HEALTH CARE

Alan M. Zaslavsky

Department of Health Care Policy, Harvard Medical School
180 Longwood Ave, Boston, MA 02115, USA
zaslavsk@hcp.med.harvard.edu

The Consumer Assessments of Healthcare Providers and Systems (CAHPS®) program has developed surveys that allow users of healthcare in the United States to evaluate the quality of the care they experience. Among other implementations, surveys in the Medicare system for the elderly and disabled have collected approaching 2 million responses over 13 years, assessing the quality of over 400 health plans (organizations providing care and insurance). The complex structure of the data makes hierarchical modelling an appropriate analytic tool. After describing the CAHPS surveys and the analytic methods used in standard reports, we review research using multilevel modelling strategies that address various aspects of the structure of the CAHPS data. The first fits a 2-level Fay-Herriott-type Bayesian hierarchical model to data aggregated by plan to estimate plan-level correlations among summary scores on different items. We applied exploratory factor analysis to draws from the posterior distribution of the covariance matrix and thereby make inferences about the plan-level factor structure of quality. By forming separate measures for healthier and sicker members of each plan, we were able to determine which items measured distinct dimensions of quality depending on health status. Analysis of the covariance structure of the coefficients of multiple predictors of quality provides further insight into the performance of the various health plans for patients with various characteristics. Another set of analyses evaluates the relative contributions of geography and organizational units to the various quality measures, and the amount of variation over time in each. Geographical variation predominated for aspects of member experiences that are not typically under the direct control of health plans, and the geographical effects were very stable over time. The methods discussed have potential application to description and reporting of multiple measures across other kinds of multilevel structures.

References:
O'Malley AJ, Zaslavsky AM. Domain-level covariance analysis for multilevel survey data with structured nonresponse. J Am Stat Assoc 2008;103(484):1405-1418.
Zaslavsky AM, Cleary PD. Dimensions of plan performance for sick and healthy members on the Consumer Assessments of Health Plans Study 2.0 survey. Med Care 2002;40(10):951-964.
Keenan P, Landon BE, Cleary PD, Zaborski LS, Zaslavsky AM. Geographic area variations in the Medicare health plan era. Med Care 2010;48(3):260-266.

***Alan M. Zaslavsky**, Ph.D., is Professor of Health Care Policy (Statistics) in the Department of Health Care Policy at Harvard Medical School. Dr. Zaslavsky's statistical research interests include surveys, census methodology, small-area estimation, official statistics, missing data, hierarchical modelling, and applied Bayesian methodology. His research topics in health care policy center on measurement of the quality of care provided by health plans through consumer assessments and clinical and administrative data. Other research interests include cancer services, psychiatric epidemiology, measurement of disparities in health care and effects of uninsurance. He is a member of the Committee on National Statistics (CNSTAT) of the National Academy of Sciences and has served on CNSTAT panels on census methodology, small area estimation and race/ethnicity measurement, as well as Institute of Medicine committees on measurement and reporting of health and of health care quality. He is a Fellow of the American Statistical Association.*

YOUNG STATISTICIANS SESSION - WHERE CAN STATISTICS TAKE YOU? THE EXPERIENCES OF SIX SUCCESSFUL STATISTICIANS

The aim of the Young Statisticians' Session is to encourage, motivate and inspire statisticians who are young in their statistical career, either studying or in their early years of working. Six experienced statisticians at various stages of their career will be presenting tales and experiences from their own careers to demonstrate the wide and interesting array of opportunities available to statisticians. Hear about statistical research, academic and commercial consulting, pharmaceutical statistics, government statistics and campaigning for statistical evidence to be used as the basis for public policy.

Everyone, and in particular, all young statisticians, are welcome and encouraged to attend the session and hear from Denise Lievesley, Susan Holmes, Sue Finch, Nicola Armstrong, Alex Maund and Kevin Wang. Don't miss the opportunity to come along, listen and ask questions of your presenters regarding their experiences and experiences that could be open to you!

STATISTICAL MODEL TO ESTIMATE THE EFFECT OF SMOKING ON LUNG CANCER

Moustafa Galal Moustafa[1], Ahmed Foaad Attia[2], Dr. Medhat Mohamed Abdelaal[3], Hisham Abdel-Tawab Mahran Morsy[4]

[1]Faculty of commerce- Ain shams university, Cairo, Egypt
mostafa-galal@hotmail.com

[2]Institute of Statistical Studies & Research - Cairo University, Cairo, Egypt
afoaad@hotmail.com

[3]Faculty of commerce- Ain shams university, Cairo, Egypt
medhatal@hotmail.com

[4]Faculty of commerce- Ain shams university, Cairo, Egypt
hisham-abdeltawab@hotmail.com

There is a state of uncertainty about the association between smoking and lung cancer; therefore, many studies have been independently conducted in many countries to examine this association. The objective of the present study is to perform a meta analysis to combine the results of those studies to build one opinion about this association. 36 studies that address this association were combined by using a random effect model through 4 levels. The first level is studying the association without taking into account the type of smoker, location of study, and type of lung cancer. The second level is studying the association taking into consideration the kind of smoker. The third level is studying the impact of location of the medical study in five different continents on the association. The fourth level is studying the effect of the type of lung cancer. The present study found significant association between smoking and lung cancer with 95% confidence level.

References

Egger, M., Smith, D.G., and Altman, D.G. (2001). Systematic Reviews in Health Care: Meta Analysis in Context. BMJ Publishing Group, 2nd edition.

Macaskill, P., Walter, S.D., and Irwig, L. (2001). Statistics in Medicine: A Comparison of Methods Detect Publication Bias in Meta Analysis. John Wiley & Sons, Volume 20 Issue.

Whitehead, A., (2002). Meta Analysis of Controlled Clinical Trials. John Wiley & Sons.

Hisham Abdel-Tawab Mahran *I am currently a teacher assistant in Faculty of Commerce, Ain Shams University, Cairo, Egypt. I am student in the preliminary study for PHD of Applied Statistics in academic year 2010/2011 in the same college.*

REVEALED PREFERENCE MODELS FOR NETWORK INFERENCE

Ryan Admiraal[1], *Mark S. Handcock*[2]

[1] *Murdoch University, 90 South Street, Murdoch, WA, 6009, Australia*
R.Admiraal@murdoch.edu.au

[2] *University of California at Los Angeles, Los Angeles, CA, 90095, USA*
handcock@ucla.edu

Networks have commonly been used to represent relational data, but, historically, most research has focused on descriptive measures of networks and not network inference. Only recently, the development of exponential random graph models (ERGMs), or p* models, has led to a viable class of stochastic models for network inference. These models, first proposed as the dyad independence p1 model by Holland Leinhardt (1981) before being extended to allow for dyad dependence, provide a unified framework for not only network inference but also network simulation.

We describe a new class of models being developed for network inference. This class of models, which we call revealed preference models (RPMs), has its origins in the economic theory of stability of two-sided markets, which attempts to describe the set of conditions under which a proposed matching consisting of pairs of agents on opposite sides of a market (e.g., universities and students) will be adhered to by all possible coalitions of agents (Roth and Sotomayor, 1990). RPMs were developed by Logan, Hoff, and Newton (2008) for one-to-one matchings and work under the assumption that an observed matching is stable. This matching can then be used to estimate agents' preferences for certain characteristics. For instance, Logan et al. use data for single individuals and married couples to estimate male and female preferences for partners of various age, education, and religion categories. We extend this work to allow for many-to-one and many-to-many matchings, specifically considering this in the context of heterosexual partnership networks where individuals may have multiple partners at a given time.

References

Holland, P.W. and Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs (with comments by R.L. Breiger, S.E. Fienberg, S.S. Wasserman, O. Frank and S.J. Haberman and a reply by the authors). Journal of the American Statistical Association, 76 (373), 33-65.

Logan, J.A., Hoff, P.D., and Newton, M.A. (2008). Two-sided estimation of mate preferences for similarities in age, education, and religion. Journal of the American Statistical Association, 103 (482), 559-569.

Roth, A.E. and Sotomayor, M.A.O. (1990). Two-Sided Matching: A Study in Game-Theoretic Modelling and Analysis. Cambridge University Press.

Ryan Admiraal *currently works as a lecturer at Murdoch University, Perth. His primary area of research interest is social network analysis, speci_cally the use of exponential family random graph models, revealed preference models, and sequential importance sampling in inference for heterosexual partnership networks.*

ANALYSIS OF SHORT INTERRUPTED TIME SERIES USING A MARGINAL LIKELIHOOD METHOD

Muhammad Akram[1],Andrew Forbes[2],Catherine Forbes[3]

[12] Department of Epidemiology and Preventive Medicine
Monash University, The Alfred Centre
99 Commercial Road, Melbourne VIC 3004.
[1]Muhammad.Akram@monash.edu
[2] Andrew.Forbes@monash.edu

[3] Department of Econometrics and Business Statistics
Monash University, Clayton campus, VIC.
Catherine.Forbes@buseco.monash.edu.au

Interrupted time series designs arise often in the evaluation of population health intervention programs, such as mass media campaigns. The data consists of the repeated observation of a variable in the population before and after a population level intervention, such as in a mass media campaign to promote HIV testing. These time series are often very short in length, and as such they pose challenges to the use of routine statistical methods for time series analysis, largely due to the poor estimation of the required autocorrelation parameters with these methods.

In this paper, we consider an AR(1) regression model for short interrupted time series'. Prior work (McKnight at al 2000) has proposed a double application of the bootstrap in which bias correction of a standard residual- based estimator of the autocorrelation parameter is performed in the first application and variance estimation for regression model parameter estimators in the second. Here we propose and evaluate an alternate approach using maximum marginal likelihood for estimation of the autocorrelation parameter not previously applied in the interrupted time series literature. Using Monte Carlo simulations we compare the performance of regression model parameter estimators using the maximum marginal likelihood with that of generalized least squares based methods (eg Prais-Winsten), both with and without the double application of the bootstrap. Our results indicate that the performance (bias, size, power, confidence interval coverage, mean squared error) of the maximum marginal likelihood estimation approach without any bias correction being applied matches or exceeds that of double-bootstrapped approaches. Furthermore, applying the double bootstrap to the maximum marginal likelihood estimator offers no additional benefit. This finding has the potential to enable faster and more efficient analyses of short interrupted time series' as well as providing an opportunity for a detailed study of design aspects of such series'.

References

McKnight, S., McKean, J. and Huitema, B.
(2000). A double bootstrap method to analyze linear models with autoregressive error terms. Psychological Methods, 5, 87-101.

Muhammad Akram *currently works as a Research-Fellow in the Department of Epidemiology & Preventive Medicine, School of Public Health & Preventive Medicine, Monash University in Melbourne. His area of interest is analysis of interrupted time series studies in health care research. Akram has also been involved in air pollution and Cardiac arrest study. His other fields of interest are Econometric modelling, Forecasting and exponential smoothing.*

MARGINAL LONGITUDINAL CURVES ESTIMATED VIA BAYESIAN PENALIZED SPLINES FOR UNBALANCED CASE

Al Kadiri M. A.[1], *Bani-Mustafa A. S.* [1], *Finch C. F.* [2]

[1] *Graduate School of Information Technology and Mathematical Sciences, University of Ballarat , Ballarat 3350, VIC, Australia*
m.alkadiri@ballarat.edu.au

[2] *School of Human Movement and Sport Sciences, University of Ballarat Ballarat 3350, VIC, Australia*
c.finch@ballarat.edu.au

This article proposes a penalized regression spline approach to represent marginal longitudinal models based on a semiparametric mixed models framework. The most practical unbalanced case, in which data is missing or different number of measurements for a set of subjects, is proposed here. A Bayesian Markov chain Monte Carlo inference with the Gibbs sampler is used to estimate model parameters. The R programming software is used and special codes written since it is not possible to fit similar models using existing codes like lme(). As an example, the Six Cites Air Pollution data is used to estimate the marginal curve of a function describing lung growth for set of children in an unbalanced longitudinal study.

References

Al Kadiri, M., Carroll, R., and Wand, M. (2010). Marginal Longitudinal Semiparametric regression via Penalized Spline. Statistics and probability letters, (accepted).
Welsh, A., Lin, X., and Carroll, R. (2002). Marginal Longitudinal Non-parametric Regression: Locality and E_ciency of Spline and Kernel Methods. American Statistical Association, 97, 482-493.
Zeger, S. and Diggle, P. (1994). Semiparametric Models for Longitudinal Data with Application to CD4 Cell Numbers in HIV Seroconverters.
Biometrics, 50, 3, 689-699.

Mohammad Al Kadiri *finished his 2nd year as a PhD student in Graduate School of Information Technology and Mathematical Sciences at University of Ballarat. His area of study is the Generalized Linear Mixed Models (GLMM)s and its applications in different practical fields. Recently, he published and submitted articles related to this area of research with applications to longitudinal/cluster data for the LMM case. More specifically, he investigated penalized regression splines as a smoothing technique to fit unknown relationships between regression variables. Implementations for different models, like additive and varying coefficient models, were studied. Currently, Al Kadiri is working to develop smoothing techniques for longitudinal data in their GLMM extensions. Some articles will be submitted soon. From his masters degree, Al Kadiri published two articles on sampling techniques, particularly, on the ranked set sampling technique. He has worked for more than 8 years as a lecturer in two universities teaching statistical courses for undergraduate students.*

RECENT DEVELOPMENTS IN INDIRECT TREATMENT COMPARISONS

Demissie Alemayehu

Pfizer and Columbia University
235 East 42nd Street - 219-8057, New York, NY 10017
alem@stat.columbia.edu

When a direct assessment of the comparative benefits of alternative treatment options is not available, indirect treatment comparisons have been proposed as a viable approach in comparative effectiveness research. This is particularly of interest in pricing and re-imbursement reviews where the only available data from placebo-controlled studies are insufficient to make recommendations on the relative risk-benefits of available treatment options. A valid use of the available techniques, however, requires stringent assumptions, which are not adequately studied in the literature. We review recent developments in the field, discuss the issues associated with the commonly used methods, propose potential remedial measures, and suggest relevant topics for further research.

Demissie Alemayehu *currently works as an Executive Director of Statistics with Pfizer, Inc., and as an Adjunct professor of Statistics at Columbia University, both in the City of New York. His area of interest is application of statistics in the design and analysis of clinical trials, with further emphasis on patient reported outcomes and comparative effectiveness research. Demissie is a Fellow of the American Statistical Association, and has extensive pharmaceutical, research and teaching experiences.*

SAMPLING MONITORING AND ADJUSTMENTS FOR INDIGENOUS SURVEYS

Tamie Anakotta[1], *Geoffrey Brent*[2]

[1] *Australian Bureau of Statistics, Methodology and Data Management Division*
Locked Bag 10, Belconnen ACT 2616
tamie.anakotta@abs.gov.au

[2] *Australian Bureau of Statistics, Methodology and Data Management Division*
GPO Box 2796Y, Melbourne VIC 3001
geoffrey.brent@abs.gov.au

Sampling a rare population efficiently can be challenging. Indigenous Australians (Aboriginal and Torres Strait Islanders) occupy a unique and important place in Australian society and culture; hence an important part of the Australian Bureau of Statistics' (ABS) objectives is to collect quality results about Indigenous persons. However, due to the complexities in locating and identifying Indigenous Australians it is difficult to design a cost-effective survey which will guarantee quality results. High migration rates of Indigenous Australians, their willingness to identify as Indigenous and their willingness to respond are some of the factors which jeopardise the quality of Indigenous survey outputs. These uncertainties can cause a high volatility in achieved sample takes, potentially leading to cost inefficiencies and insufficient quality of estimates. This presentation focuses on how the ABS monitored the sample take of its recent Indigenous social survey in light of increases in management information from the field interviewers. Analyses of weekly updates on response rates and sample loss rates were used to adapt the design, during enumeration, to improve the quality of the results.

References

Brent, G., Rogers, A. (2008). Sample Design Issues for National Surveys of the Indigenous Population. Australian Bureau of Statistics, 1352.0.55.096

Groves, R.M., Heeringa, S.G. (2006). Responsive design for household surveys: tools for actively controlling survey errors and costs. Journal of Royal Statistical Society A, 169, Part 3, pp. 429-257

Tamie Anakotta *currently works in the Household Methodology Unit at the Australian Bureau of Statistics. She provides methodological support for the sample design, sample monitoring and estimation for various household surveys in particular Indigenous surveys.*

ACCOUNTING FOR UNCERTAINTY IN EXTREMAL DEPENDENCE MODELLING USING BAYESIAN MODEL AVERAGING TECHNIQUES

P.Apputhurai and A.G.Stephenson

Psychological Sciences and Statistics, Faculty of Life and Social Sciences
Swinburne University of Technology, VIC 3122, Australia
papputhurai@swin.edu.au

Modelling the joint tail of an unknown multivariate distribution can be characterized as modelling the tail of each marginal distribution and modelling the dependence structure between the margins. Classical methods for modelling multivariate extremes are based on the class of multivariate extreme value distributions. However, such distributions do not allow for the possibility of dependence at finite levels that vanishes in the limit. Alternative models have been developed that account for this asymptotic independence, but inferential statistical procedures seeking to combine the classes of asymptotically dependent and asymptotically independent models have been of limited use. We overcome these difficulties by employing Bayesian model averaging to account for both types of asymptotic behaviour, and for subclasses within the asymptotically independent framework. Our approach also allows for the calculation of posterior probabilities of different classes of models, allowing for direct comparison between them. We demonstrate the use of joint tail models based on our broader methodology using two oceanographic datasets and a brief simulation study.

Pragalathan Apputhurai *is a Ph.D. student of Statistical Science at Hawthorn campus, Swinburne University of Technology, Australia. He received the B.Sc degree and the M.Sc degree from University of Peradeniya, Sri Lanka. His research interests include spatial statistics, modelling extreme values and complex modelling of environmental data.*

REGULARIZED FUNCTIONAL CLASSIFICATION OF THE KINETIC TRACE OF THE DEVELOPMENTAL MOUSE RETINA

Yuko Araki[1], Takashi Yanagawa[2]

[1,2]Biostatistic Center, Kurume University
67 Asahi-machi, Kurume, 830-0011, Japan
[1] Araki_yuuko@med.kurume-u.ac.jp, [2] yanagawa_takashi@kurume-u.ac.jp

For the investigation of a complex phenomenon underlying biological transformation and transition such as the complex molecular events underlying the postnatal development of the mouse retina, kinetic trace data of protein expression play a critical role (Haniu et al. 2006). Although many classification analyses have been performed for protein expression data using statistical methods such as hierarchical clustering, self-organizing maps or support vector machines, developing adequate models to analyze time course data is urgent. To incorporate the information that is inherent in processes over time, we propose adopting a method of classifying a set of traces or curves in which individual traces are modelled as realizations of smooth functions of time.

The method uses a recently developed functional classification tool (Araki et al. 2009) based on Gaussian radial basis expansions with the help of regularization and functional logistic discrimination with a model selection technique, generalized Bayesian information criterion. The discrimination tool is designed to construct a decision rule based on data given as a set of functions. The model was used to classify the proteomic trajectory of the postnatal development of the mouse retina into known groups. The method is applicable to classify any biological transformation and will be a powerful tool in biomedical sciences.

References

Araki, Y., Konishi, S., Kawano, S., and Matsui, H., 2009. Functional logistic discrimination via regularized basis expansions. Communications in Statistics-Theory and Methods, 38, pp.2944-57.
Haniu, H., Komori, N., Takemori, N., Singh, A., Ash, J.D. and Matsumoto, H.,2006. Proteomic trajectory mapping of biological transformation: Application to developmental mouse retina. Proteomics, 6, pp.3251-61.

Yuko Araki *is an Assistant Professor at the Biostatistics Center, Medical School, Kurume University , Japan. She earned her Ph.D at Kyushu University, Japan. Her interests include statistical modelling for longitudinal data in medical/biological science, causal modelling for survival data, and developing information criterion for model selection. She teaches the first course of biostatistics to graduate students at the Biostatistics Center, Kurume University, and also works as a statistical consultant for medical doctors at Kurume University hospital. Currently she has been involved in some research projects;*
-Epidemiological studies of atomic bomb survivors in Japan,
-An efficient procedure development for model selection with information criterion,
-Modelling/applying functional classification methods to the biological transformation data of the developmental mouse retina.

AN AUSTRALIAN RISK PREDICTION MODEL FOR DETERMINING EARLY MORTALITY FOLLOWING AORTIC VALVE REPLACEMENT

Thathya V. Ariyaratne[1], Baki Billah[1], Cheng-Hon Yap[1], Diem Dinh[1], Christopher M. Reid[1]

[1]Department of Epidemiology and Preventive Medicine, 6th floor, The Alfred Centre
99 Commercial Road, Melbourne, VIC 3004, Australia
thathya.ariyaratne@med.monash.edu.au

Objective: To identify risk factors associated with early mortality following aortic valve replacement (AVR) in Australian patients, and to develop a multivariable logistic model for pre-operative risk prediction.

Methods: Prospectively collected data from the Australasian Society for Cardiac and Thoracic Surgeons (ASCTS) database project was used. AVR procedures performed between July 2001 and June 2008 were included for analysis. Preoperative variables with a p-value of < 0.10 in chi-squared analysis were considered for multiple logistic regression analysis. Using a bootstrap re-sampling technique, five candidate models were identified. Models were validated internally using average receiver operating characteristic (ROC) curves and Hosmer Lemeshow (H-L) goodness-of-fit tests via the bootstrap (n-fold) validation method (Hosmer, 1989 and Billah et al, 2010). In addition to model performance, the Akaike Information Criterion (AIC) (Akaike, 1970) and prediction mean square error (MSE) were considered in the selection of the final model (AVR-Score). The AVR-Score was externally validated on 1258 consecutive procedures performed between 01 July 2008 and 30 June 2009.

Results: Between July 2001 and June 2008 a total of 3544 AVR procedures were performed. Early mortality was 4.15%. The final model contained the following variables: age, New York Heart Association class, left main disease, infective endocarditis, cerebrovascular disease, renal dysfunction, previous cardiac surgery, and estimated ejection fraction. Internal validation of the AVR-Score yielded an average area under the ROC curve of 0.779 (95%CI: 0.762, 0.795) and an H-L p-value of 0.412 (p>0.05), indicating good discrimination and calibration capacity. External validation of AVR-Score produced a ROC of 0.730 (0.706, 0.755) and H-L p-value of 0.478.

Conclusion: We have identified 8 key predictors of early AVR mortality in Australian patients and developed a preoperative risk prediction model for early mortality.

References

Hosmer DW, Lemeshow S. Applied logistic regression. New York: John Wiley & Sons; 1989.
Billah B, Reid CM, Shardey GC, Smith JA. A preoperative risk prediction model for 30-day mortality following cardiac surgery in an Australian cohort. Eur J Cardiothorac Surg. 2010;37(5):1086-92.
H. Akaike, Statistical predictor identification. Annals of the Institute of Statistical Mathematics. 1970; 22(1) 203–217.

Thahtya V. Ariyaratne *is currently a PhD student at the Department of Epidemiology and Preventive Medicine (DEPM), Monash University. The title of her project is the 'Comparison of long-term outcomes and cost-effectiveness of coronary artery bypass graft and percutaneous coronary interventions ". She joined the department as an Honours student in 2009 following the completion of her bachelor of Biomedical Science/ Bachelor of Economics at Monash University. The above abstract presents the work of her Honours thesis, which was ranked second (First class; rank #2) in the cohort of Biomedical Science students completing Honours projects in 2009. At present, Thathya is a member of the academic teaching staff of DEPM involved in the wider dissemination of education in Epidemiology and Biostatistics to undergraduate students and postgraduate students from the Victorian Consortium for Public Health.*

WHEN GENOMIC DATA MEETS CLINICAL TRIALS

Nicola Armstrong

Cancer Research Program, Garvan Institute of Medical Research

Design of clinical trials is a well documented area of biostatistics. However, the era of personalised medicine is fast approaching, and a large part of genomic research is currently focused on identifying prognostic and predictive biomarkers for treatment of disease, especially for cancer. Incorporating this sort of information into a clinical trial design does not always follow the text book examples. In this talk, I will discuss some of the studies I have been involved with and the area of translational research in general.

Nicola Armstrong *received her PhD in statistics at UC Berkeley with Prof Terry Speed and spent several years at the Dutch Cancer Institute (NKI-AvL) before moving to the Garvan Institute in Sydney earlier this year. Her research focus is on genetics and genomics data.*

STATISTICAL METHODS FOR GENETIC ASSOCIATION ANALYSES IN THE RESEQUENCING ERA

David Balding

UCL Genetics Institute
2nd Floor, Kathleen Lonsdale Building
5 Gower Place, London WC1E 6BT UK
d.balding@ucl.ac.uk

The tsunami of genome-wide association studies that swept across the world in 2007-2009 has now dissipated, leaving us with many novel associations between common genetic variants and complex diseases, and many interesting leads towards causal mechanisms. But there has been no immediate great advance in understanding disease mechanisms, and only a small fraction of the heritable variation in disease-related phenotypes has been explained. Whole-exome, and eventually whole-genome, resequencing is the next big thing in the search for genetic mechanisms of disease, and will allow us to identify rare variants of relatively large effect, for which causal mechanisms are expected to be more obvious. But resequencing data brings with it a host of both new and familiar statistical problems. The first of these revolve around data quality and the measurement of confidence in base calling. The next problem is how to use gappy and variable-quality sequence data to generate honest measures of association. Remote kinship is another important factor: a group of apparently unrelated individuals may share a rare variant due to inheritance from an unsuspected common ancestor a handful of generations in the past. Remote kinship creates both risks and opportunities. The risk is of confounding in an association analysis: these individuals may share several genomic regions, and a signal at one locus may in part be caused by a causal variant at another locus. The opportunity is that the identification of remote shared ancestors can allow “population linkage" analysis which can enjoy the immunity to confounding of linkage studies while also offering locus refinement as good or better than that achievable by association studies. Ideally it may be possible to combine linkage and association analyses of resequencing datasets to achieve increased power. I will review these approaches and try to suggest avenues for future advances.

David Balding *received a B.Math. from the University of Newcastle (NSW) and a D.Phil. in mathematics (applied probability) at the University of Oxford, UK. He then held a junior academic post at Oxford for a year before moving successively to Queen Mary London, the University of Reading, and Imperial College London, where he was Professor of Statistical Genetics in the Department of Epidemiology and Public Health from 2001 to 2009. In October 2009 he joined the new Institute of Genetics at University College London. He researches a wide range of mathematical and statistical problems in genetics - evolutionary, population, forensic and medical. He is lead editor of the Handbook of Statistical Genetics (3rd edn, Wiley, 2007) and has published a monograph on the interpretation of DNA profile evidence.*
For more information see http://www.zebfontaine.eclipse.co.uk/djb.htm.

FORECASTING MODELS OF FLOOD-AFFECTED RESIDENTIAL PROPERTY PRICES

Abdul Mutalib Beksin

University of Malaya
50603 Kuala Lumpur MALAYSIA
talibkf@siswa.um.edu.my

This work is motivated by the research gap evident in the area of forecasting models for flood affected residential properties in Malaysia. The predominant focus is on an empirical investigation of several transacted price forecasting models of a city on the east coast of Malaysia. Their applicability and performance are analyzed and city as well as forecasting horizon specific patterns are determined and interpreted.

After the literature review, mostly on Anglo-Saxon research, I derive the theoretical foundations which are important in executing the empirical part of the work. Therefore, I discuss theoretically general real estate market characteristics, the specifics of time series and panel data, common forecasting models, and forecasting techniques as well as performance measures.

The major finding of the empirical work, which contains the transacted price series investigation, is that ARIMA and multivariate regression models are generally able to forecast price series for the flood affected residential properties market. Furthermore, I observed that single ARIMA models perform well for forecasting horizon of less than two years. Moreover, univariate models outperform multivariate regression models in the short run. On the other hand, multivariate regression models outperform the univariate models in the longer run.

Precise forecasts require starting values at economically reasonable price levels. The major finding of the second part of the empirical work is the general ability of multivariate regression models to forecast price series and prudently constructed models outperform more complex models for the short run and for the long run.

Abdul Mutalib Beksin *- In between juggling multiple research projects on forecasting, the interface between environment and real estate, project management and casual teaching at the University Of Malaya in Kuala Lumpur, Mutalib hunts high and low for conferences or seminars especially in Western Australia; not for his insatiable hunger for knowledge but more so for his love of seafood, especially raw, fresh oyster.*

GROUPED SUBSET SELECTION

Yi Guo[1], Mark Berman[2], Andy Green[3]

[1] CSIRO Mathematics, Informatics and Statistics
Locked Bag 17, North Ryde NSW 1670
yi.guo@csiro.au

[2] CSIRO Mathematics, Informatics and Statistics
Locked Bag 17, North Ryde NSW 1670
mark.berman@csiro.au

[3] OTBC Pty. Ltd.
8 Lawley Crescent, Pymble NSW 2073
andy.green@ozemail.com.au

The work described here is motivated by a problem in spectroscopy. Over the years, we have built a library of about 500 spectra, each measured at about 300 wavelengths, and representing 60 "pure" (mostly mineral) classes. So on average, there are about 8 samples per pure class. Real world spectra often contain mixtures of 1, 2, 3 or 4 minerals. We "unmix" spectra of such minerals using a relatively simple linear mixture model (with non-negative weights), a regularised version of the average 300 x 300 within-class covariance matrix based on our library of pure spectra, and fast subset selection procedures using code written by Alan Miller (Miller, 2002). The code is regularly used in the mining industry to unmix tens of thousands of spectra measured on drill core samples. So speed is an important consideration.

An examination of the pure spectra in our library reveals that different classes have different within-class covariance matrices, so that our (implicit) assumption of a common within-class covariance matrix is incorrect. This means that sometimes classical statistical inference procedures fail, especially when there are 3 or 4 materials in the mixture. To reduce this problem (while still producing reasonably fast solutions), we have started modelling each material in our library as itself a mixture of 2 (or more) materials, rather than by using the class mean. The implication of this is that, when applying subset selection to mixed spectra, we need to take subsets of pairs (or more generally groups) of spectra. So, for instance, modelling a spectrum as a mixture of 2 materials, means modelling it as a mixture of 2 pairs of 2 spectra. Unfortunately, Miller's code does not obtain best fitting subsets of groups of classes.

We have developed fast software that carries out grouped subset selection. The underlying algorithms will be described and its application to some spectra illustrated.

Reference

Miller, A. (2002**).** *Subset Selection in Regression,* Second edition. London: Chapman and Hall.

Mark Berman *is a research scientist with CSIRO Mathematics, Informatics and Statistics, Sydney. His research interests are in image analysis (especially hyperspectral), spectroscopy and spatial data analysis. In recent years, he has been working primarily on the mixture analysis of large volumes of spectra and airborne hyperspectral images focused on exploration and environmental applications.*

DERIVING TESTS OF THE SEMI-LINEAR REGRESSION MODEL USING THE DENSITY FUNCTION OF A MAXIMAL INVARIANT

Jahar L. Bhowmik [1], Maxwell L. King [2]

[1] Faculty of Life and Social Sciences, Swinburne University of Technology
John Street, Hawthorn, VIC 3122, Australia
jbhowmik@swin.edu.au

[2]Department of Econometrics and Business Statistics, Monash University
Clayton, VIC 3800, Australia
Max.King@adm.monash.edu.au

In the context of a general regression model in which some regression coefficients are of interest and others are purely nuisance parameters, we derive the density function of a maximal invariant statistic with the aim of testing for the inclusion of regressors (either linear or non-linear) in linear or semi-linear models. This allows the construction of the locally best invariant test, which in two important cases is equivalent to the one-sided t test for a regression coefficient in an artificial linear regression model.

References

Bhowmik, J.L, and King, M.L. (2007). 'Maximal invariant likelihood based testing of semi-linear models', Statistical Papers, vol.48, pp. 357-383.
Bhowmik, J.L, and King, M.L. (2009). 'Parameter estimation in semi-linear models using a maximal invariant likelihood function', Journal of Statistical Planning and Inference, vol.139, pp.1276-1285.
King, M.L. (1980). 'Robust tests for spherical symmetry and their application to least squares regression', Annals of Statistics, vol. 8, pp.1265-1271.

Jahar Bhowmik *currently works as a lecturer in statistics at the Faculty of Life and Social Sciences of Swinburne University in Melbourne. He received a PhD degree in Applied Statistics and Econometrics from Monash University in 2004. Jahar worked for many universities including Monash University and Tasmania University. His research interests are Statistical Inference (linear and non-linear models), Biostatistical Data Modelling (trend analysis in cancers, BMI), Maximal Invariant Likelihood based Inference (non-linear models), and Experimental Design. Jahar has been involved in statistical consulting, teaching statistics, and postgraduate students' supervision.*

MODELLING SURVEY RESPONDENT BEHAVIOUR: A SURVIVAL ANALYSIS APPROACH

Melanie Black

Australian Bureau of Statistics
45 Benjamin Way Belconnen ACT 2616
melanie.black@abs.gov.au

ABS businesses surveys are mostly mail-out, mail-back surveys with telephone and reminder letter follow-up to encourage providers to respond. This follow-up is expensive and time consuming, and providers can respond at any time during the enumeration period (and not only during a call made by the ABS). Therefore, the impact of various follow-up activities on response behaviour is not always obvious or easily established. In addition, follow-up strategies are often strongly related to respondent characteristics (including historical respondent behaviour) and therefore confounding is an issue when investigating the impact of a strategy. Survival analysis (with strategies nested within respondent characteristics) is used to understand how response behaviours relate to time, as well as allowing the assessment of the effectiveness of various follow-up activities such as how long to wait between sending reminder letters, and which businesses are most likely to respond if given an earlier reminder call. Response rates can then be predicted ahead of time, based on specified follow-up strategies for different types of providers. When combined with models for forecasting costs, this provides a powerful tool for choosing an appropriate, cost-effective follow-up strategy.

Melanie Black *currently works in the Operations Research and Process Improvement team at the Australian Bureau of Statistics, where she identifies and implements improvements to the cost effectiveness and efficiency of ABS business survey operations, particularly contact and follow-up strategies, through the analysis of paradata.*

ESTIMATION OF A COMMON MODEL FOR REPLICATED TIME SERIES REALISATIONS

Ross Bowden and Brenton R. Clarke

Mathematics and Statistics, Murdoch University
South St, Murdoch, Western Australia, 6970
ross.bowden@iinet.net.au

This talk presents a method for fitting a single ARMA time series model simultaneously to multiple independent realisations of a time series process. It shows that the natural time series representations of this problem cannot be fitted efficiently using readily available software. However by first interleaving the time series it is possible to employ existing univariate modelling tools. The interleaving approach and its properties will be presented along with its application to other time series models and to daily maximum temperatures.

Ross Bowden *currently works as a Statistical and Pricing Consultant in Perth and is also undertaking a PhD at Murdoch University. His primary area of interest is time series analysis, in particular, when applied to solar radiation measurements. Ross has 35 years experience in the energy utility industry and has undertaken projects in forecasting, energy use research, market research and competitive pricing.*

MODELLING OPERATIONAL COSTS IN THE MONTHLY POPULATION SURVEY

Geoffrey Brent

Operations Research and Process Improvement Unit, Australian Bureau of Statistics
485 Latrobe Street, Melbourne, Vic 3000
geoffrey.brent@abs.gov.au

Face-to-face interviewing work in the ABS' Monthly Population Survey involves a mix of highly clustered and less-clustered households, and decisions on the best contact procedure require a fine-level understanding of how effort spent on these different households contributes to survey costs and outcomes. Cost data is recorded at the level of trips, which may include a mix of different types of work and some activities relating to multiple households (e.g. travel to/from the region). In order to get a better breakdown of how fine-level activities contribute to costs, we use a combination of simple regression models. The choice of models is guided not only by predictive accuracy on existing data, but also by the nature of the decisions that the model is expected to inform, and the need to integrate with other models for the survey response rate and the survey estimation bias. When integrated with a response model, this approach reveals that some activities are considerably more effective than others (in terms of responses gained per dollar spent), and shifting effort towards these activities may improve overall survey outcomes.

Geoffrey Brent *joined the ABS three years ago, after nine years in biomedical engineering. He started in household survey methodology, working on a targeted sample design for the 2008/9 National Aboriginal and Torres Strait Islander Social Survey, and subsequently moved to the Operations Research branch in Melbourne.*

STATISTICAL METHODOLOGIES IN THE PHARMACEUTICAL INDUSTRY: APPLICATIONS TO COMMERCIALISATION IN ELI LILLY

Alan J M Brnabic

Intercontinental Information Sciences, Eli Lilly Australia Pty Ltd
Level 1, 16 Giffnock Avenue, Macquarie Park, NSW 2113, Australia
Brnabic_alan@lilly.com

Working for Eli Lilly as a statistician presents many opportunities to develop and apply statistical methodologies to attempt to address the plethora of questions related to data collected in the pharmaceutical industry. Statisticians can work across any phase of research and can be involved in concept development, design, analysis, programming, reporting, interpretation and presentation of this research. With the ease of access to more sophisticated methods such as data mining, Bayesian statistics, adaptive designs and tools for comparative effectiveness Eli Lilly is leveraging both internal and external expertise to help solve increasingly complex problems which may not be best answered from a randomised controlled trial (RCT). With growing interest in the use of observational, real-world data to answer questions for the patient, payer and prescriber, methods such as propensity scoring, matching and marginal structural models (MSM) are useful additions to the analyst's toolkit. Propensity scoring and matching have been used widely to attempt to address the issue of treatment allocation bias, however the choice of which approach to take is not always clear. Issues arising from switching treatments during prospective observational studies can be modelled using MSM by incorporating treatment as a time-varying covariate. In conclusion, a statistician in the pharmaceutical industry now faces many challenges beyond the realm of the traditional RCT, These often require new methodologies along with an appraisal of when it is appropriate to use them. This is both challenging and exciting.

References

Baser O (2006). Too much ado about propensity score models? Comparing methods of propensity score matching. *Value in Health, 9*(6), 377-85.

Hansen BB (2008). The essential role of balance tests in propensity-matched observational studies: Comments on 'A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003' by Peter Austin, *Statistics in Medicine*. *Statistics in Medicine, 27*, 2050-2054. Retrieved from http://www.epi.msu.edu/janthony/requests/propensity/Hansen_Commentary_1.pdf

Hernán, M.A., Brumback, B., Robins, J.M. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men*. Epidemiology 11*, 561-570.

Mr Brnabic *is the Health Outcomes and Data Mining Statistics Manager/research scientist at Eli Lilly Intercontinental Information Sciences. His work includes: leading and managing staff, designing/reviewing concepts and studies (Phase IIIb & IV observational studies), analyzing/reviewing studies, presenting & writing/reviewing results in peer review publications, leading and reviewing external methodologies/ guidelines for use within the company as well as consulting/coordinating strategy for analysis on Reimbursement dossiers & other related Health Outcome activities for countries like Australia, Canada & Korea. Mr. Brnabic received a BA Dip Ed in mathematics and statistics and a MA in applied statistics from Macquarie University, Sydney. Mr. Brnabic's interests are in the design and analysis of observational studies, and practical applications of statistical methodology to the field of Neuroscience. He is also interested in Health Outcomes and its role in reimbursement of medicines.He has A-STAT Professional Accreditation with the Statistical Society of Australia (SSAI). He is an active member and co-chair for the Australian Pharmaceutical Biostatistics Group (APBG).*

SPATIAL POINT PATTERN ANALYSIS TO INVESTIGATE ECOLOGICAL PROCESSES

C Brown[1], J Illian[2], D Burslem[2], R Law[4]

[1] University of St Andrews, Centre for Research into Ecological and Environmental Modelling
The Observatory, Buchanan Gardens, St Andrews, Fife, KY16 9lZ, UK
calum@mcs.st-and.ac.uk

[2] University of St Andrews, Centre for Research into Ecological and Environmental Modelling
The Observatory, Buchanan Gardens, St Andrews, Fife, KY16 9lZ, UK
janine@mcs.st-and.ac.uk

[3] University of Aberdeen, School of Biological Sciences
Zoology Building, Tillydrone Avenue, Aberdeen, AB24 2TZ, UK
d.burslem@abdn.ac.uk

[4] University of York, Department of Biology (Area 4)
PO Box 373, York, YO10 5YW, UK
rl1@york@ac.uk

A variety of contrasting mechanisms have been proposed to explain observed structure, relative abundance patterns and species coexistence in biodiverse ecological communities. However, it has proved difficult to distinguish their effects and hence to assess their validity on the basis of first-order properties such as species diversity and abundance distributions (McGill et al., 2007).

We assess the relatively neglected second-order spatial implications of these mechanisms. Not only may these be quite distinct and allow the identification of specific ecological processes, but they can also be compared to spatially-explicit data already available for a range of plant communities (e.g Wiegand et al., 2007).

A stochastic individual-based model in continuous space is developed to generate multi-species spatial patterns under contrasting theories of coexistence. These are then analysed using spatial point process statistics in an attempt to identify the spatial signals of each theorised ecological process. First-order characteristics are also investigated, and the abilities of the spatial and non-spatial measures to distinguish underlying processes are compared.

These methods are subsequently applied to tropical rainforest data in order to determine the relative merits of different statistics in describing spatial structure. The potential for empirically discerning specific processes is therefore assessed and those consistent with observed patterns are identified.

References

McGill, B. J., Etienne, R. S., Gray, J. S., Alonso, D., Anderson, M. J., Benecha, H. K., et al. (2007). Species abundance distributions: Moving beyond single prediction theories to integration within an ecological framework. Ecology Letters, 10(10), 995-1015.
Wiegand, T., Gunatilleke, S., & Gunatilleke, S. (2007). Species associations in a hetergeneous Sri Lankan dipterocarp forest. The American Naturalist, 170(4), E77-E95.

Calum Brown *is a PhD student in statistical ecology at the Centre for Research into Ecological and Environmental Modelling at the University of St Andrews. His research focuses on the use of spatial statistics in identifying ecological processes that generate and maintain biodiversity. His first degree was in Physics and he has an MSc in Ecology.*

THE GENERALISED RANDOM TESSELLATION STRATIFICATION APPROACH TO DESIGNING STREAM HEALTH MONITORING SURVEYS IN QUEENSLAND

Burridge, Charis Y[1], Dobbie, Melissa J[2], Stevens, Don L, Jr[3]

[1] CSIRO, Mathematics, Informatics and Statistics Division, CSIRO Marine Laboratories
233 Middle Street, PO Box 120, Cleveland, Qld 4163, Australia
Charis.Burridge@csiro.au

[2] CSIRO, Mathematics, Informatics and Statistics Division,
Building 29, Long Pocket Laboratories, 120 Meiers Rd, Indooroopilly, QLD 4068, Australia
Melissa.Dobbie@csiro.au

[3] Department of Statistics, Oregon State University
Corvallis, Oregon, USA 97331
stevens@science.oregonstate.edu

Ecological surveys such as stream network monitoring have an inherently spatial structure, and indicators tend to exhibit spatial correlation so that it is desirable to spread sites out over the landscape in a semi-systematic fashion. Furthermore, a variety of practical problems is often encountered when running such surveys: inaccuracy of maps of the stream network, difficulty in physically accessing selected sample sites or obtaining permission to sample there (leading to non-response). Furthermore, there may be a need to stratify on features, such as stream order, that do not partition the landscape into convenient spatially-contiguous sub-regions. Dobbie and Henderson (2008) compared a variety of approaches to designing such surveys. One that stands out is Generalised Random Tessellation Stratification (GRTS), a flexible survey design method developed by Stevens and Olsen (2004) that addresses many of the issues encountered in designing and implementing ecosystem monitoring surveys. Stevens and Jensen (2007) also discuss the efficiency of various designs, including GRTS, for wetland surveys. We will describe how we applied the GRTS approach to stream health surveys in Queensland. We also introduce a random-clustering variant of ours that reduced travel time when a high proportion of selected sites could not be sampled or did not meet criteria for inclusion as reference sites.

References

Stevens, D.L., Jr. and Olsen, A.R. (2004). Spatially-balanced sampling of natural resources. Journal of the American Statistical Association, Vol. 99, pp. 262-77.

Stevens, D.L., Jr. and Jensen, S.F. (2007). Sample design, execution, and analysis for wetland assessment. Wetlands, Vol. 27, pp. 515–523.

Dobbie, M.J., Henderson, B.L. and Stevens, D.L., Jr. (2008). Sparse sampling: spatial design for monitoring stream networks. Statistics Surveys, Vol. 2, pp. 113–153.

Charis Burridge *currently works as a senior research statistician in Brisbane with CSIRO Mathematics, Informatics & Statistics Division, collaborating with marine and aquatic ecologists on the design and analysis of ecological surveys and experiments. She is interested in exploring design-based approaches to aquatic surveys but also has an interest in the model-based approach such as smoothing splines for spatial interpolation in the marine context.*

MODELLING SURVIVAL OF WILD RABBITS USING TIME DEPENDENT INDIVIDUAL COVARIATES

Kym L. Butler[1], Esther Meenken[2], Hwan-Jin Yoon[3], Khageswor Giri[4] and Steve Mcphee[5]

[1]Future Farming Systems Research Division, Department of Primary Industries
600 Sneydes Road, Werribee, Victoria, 3030, Australia
kym.butler@dpi.vic.gov.au

[2]Plant & Food Research
Private Bag 4704, Lincoln, N Z
meenkene@crop.cri.nz

[3]School of Mathematical Sciences (MSI)
Australian National University
Canberra ACT 0200, Australia
Hwan-jin.yoon@anu.edu.au

[4]Future Farming Systems Research Division, Department of Primary Industries
600 Sneydes Road, Werribee, Victoria, 3030, Australia
Khageswor.giri@dpi.vic.gov.au

[5]Agricultural Technical Services and Farming Systems Research Division, DPI
177 Ballan Road, Werribee, Victoria 3030, Australia
steven.mcphee@bigpond.com

In the development of capture-recapture models, the use of individual and environmental covariates in the estimation of parameters such as survival and recapture probability is a relatively new phenomenon. Most of the available literature focuses on the use of environmental covariates that may vary with time but are constant for individual animals and individual covariates that are fixed over the time. The use of time dependent individual covariates in capture-recapture models is a complex and difficult problem and there are not many publications on this topic (Pollock 2002). The difficulty arises in incorporating these covariates in capture-recapture models because the covariate value of an individual animal, at a given capture occasion, is only observed if that individual animal is actually captured at that capture occasion. It is not possible to simply delete any animal that has missing covariate values from the analysis. For an open population Cormack-Jolly-Seber model, we purpose a sensible methodological approach to incorporate a time dependent individual covariate in survival modelling. This approach allows us to examine the survival of individual rabbits in terms of their antibody status through time. Key statistical methodology issues that need a very careful consideration while developing such an approach are addressed.

References

Pollock, K.H. (2002). 'The use of auxiliary variables in capture-recapture modelling: an overview,' Journal of Applied Statistics, vol. 29, no 1-4, pp. 85-102.

Kym Butler *currently works as a senior Biometrician at Department of Primary Industries, Victoria. Prior to joining DPI, in 1987, Kym worked in CSIRO's Division of Mathematics and Statistics, the Queensland Department of Primary Industries and taught at the University of Queensland. Kym is a committee member of the Victorian Branch of the Statistical Society of Australia. Kym has a very wide rage of expertise in the field of applied statistics.*

INTEGRATING GENE EXPRESSION AND CLINICAL DATA IN MELANOMA PROGNOSIS PREDICTIONS

Anna Campain[1], *Yee Haw Jean Yang*[2]

[1] *Mathematics and Statistics, Center of Mathematical Biology, University of Sydney*
Carslaw Building F07 Sydney, NSW
anna.campain@sydney.edu.au

[2] *Mathematics and Statistics, Center of Mathematical Biology, University of Sydney*
Carslaw Building F07 Sydney, NSW
jean.yang@sydney.edu.au

As mircoarray and other forms of high throughput and meta-data become more readily available there is a growing need to successfully integrate gene expression and other forms of clinical, pathological and mutation data in practice. Motivated through predictive studies of an aggressive metastasised melanoma we explore some of the statistical issues associated with integrating clinical data and expression data in a classification context. Melanomas are common in particular demographics of the population. Considered a relatively minor cancer when treated successfully, a significant percent of melanomas metastasise from their primary sight. Of those patients with melanomas that metastasise, about 40% go on to live cancer free, but another 40% succumb to the disease in less than 1 year.

We examine predictive models that integrate both a clinical signature, obtained through the interaction of clinical, pathological and mutation predictors as well as a molecular signature developed through the expression data. Developing a clinical signature is potentially cumbersome. Clinical data offers a diverse range of challenges that need to be overcome. For example, missing data within clinical, pathological and mutation variables are common and problematic. We examine how the proportion of missing data can affect regression coefficients and how this translates into issues with clinical data. Missing datum can be overcome through careful multiple imputation when missingness is not too great. When subgroups of particular variables, especially pathological and mutation variables, are small, care is needed to ensure the development of a stable model. In addition, we discuss issues and challenges associated with agreement between the molecular signature obtained through our expression data and other signatures within the literature. The process of combining clinical and molecular signatures is non-trivial and we explore issues associated with this including scaling and validation.

Anna Campain *is a PhD student in the School of Mathematics and Statistics at the University of Sydney. She is completing her PhD under the advisement of Dr Jean Yang and Dr Samuel Mueller. Her research interests include the application of statistics to problems in genomic and medical research.*

MAXIMUM LIKELIHOOD ESTIMATION FOR CONTINGENCY TABLES AND LOGISTIC REGRESSION WITH INCORRECTLY LINKED DATA

James O. Chipperfield, Paul Campbell

Australian Bureau of Statistics
ABS House, 45 Benjamin Way, Belconnen ACT 2617
Paul.campbell@abs.gov.au

Data linkage is the act of bringing together records, that are believed to belong to the same unit (e.g. person or business) from two or more files. It is a very common way to enhance dimensions such as time and breadth or depth of detail. Data linkage is often not an error-free process and can lead to linking a pair of records that do not belong to the same unit. There is an explosion of record linkage applications, yet there has been little work on assuring the quality of analyses using such linked files. Naively treating such a linked file as if it were linked without errors will, in general, lead to biased estimates. This paper develops a maximum likelihood approach for analysis of linked records. The estimation technique is simple and is implemented using the well-known EM algorithm. A well known method of linking records in the present context is probabilistic data linking. The paper demonstrates the effectiveness of the proposed estimators in an empirical study which uses probabilistic data linkage.

References

Fellegi, I.P. and Sunter, A.B. (1969) A theory for record linkage. Journal of the American Statistical Association, 64, 1183-1210.
Herzog, T.N., Scheuren, F.J. and Winkler, W.E. (2007), Data Quality and Record Linkage Techniques. Springer: New York.

Paul Campbell *has a background in psychology and computer science, and is currently employed in the Analytical Services Branch of the Australian Bureau of Statistics. In the past three years he has developed a research interest in probabilistic data linking (linking individuals without the use of a unique record identifier), and is working with a team of analysts who have conducted work on both how to best link data and issues surrounding the analysis of probabilistic linked data. The aim of this work has been to both assess and minimise the impact of errors in linked data on analysis.*

MULTIPLE IMPUTATION FOR MISSING DATA: THE GENIE OUT OF THE BOTTLE?

John Carlin

Murdoch Children's Research Institute & University of Melbourne
Clinical Epidemiology & Biostatistics Unit
Royal Children's Hospital, Flemington Road, Parkville VIC 3052
john.carlin@mcri.edu.au

Multiple imputation (MI) is rapidly becoming a standard approach for handling missing data problems in epidemiology and the social sciences, with automated routines now available in major statistical packages. However, although the method has much to offer when applied judiciously, it is not a panacea for missing data (Sterne et al, 2009). Challenges facing applied statisticians in assessing the value of MI in particular applications may be considered under three major headings: (i) is MI worth considering? – assessing the potential for genuine recovery of information by use of MI, (ii) how should MI be carried out? – assessing the validity of available imputation modelling approaches, and (iii) how should MI be checked once applied? – appropriate diagnostic checking of those aspects of the imputation modelling that are amenable to empirical assessment. Our recent research has focused on aspects of (i) and (ii). First, we have compared MI with analysis using complete cases only, pointing out that the former is not always less biased, depending on the reasons for the data being missing, and that even when MI is less biased gains in precision may not be great (White and Carlin, 2010). In the context of missing covariates in regression models, gain in precision for a given parameter may be approximated by the fraction of incomplete cases among the observed values (FICO) for the corresponding covariate. Other work has compared the two widely available imputation methods (imputation based on a multivariate normal assumption, and imputation using "chained equations"), with results that surprisingly favour the multivariate normal approach even when discrete variables are imputed (Lee and Carlin, 2010). Further work is needed to develop diagnostic procedures to accompany automated procedures. MI can clearly produce poor results in some situations and statisticians need to be alert to the potential misuse of this alluringly powerful tool.

References

Lee, K. J. & Carlin, J. B. (2010) Multiple imputation for missing data: Fully conditional specification versus multivariate normal imputation. American Journal of Epidemiology. 171:624-632.

Sterne, J. A. C., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M. & Carpenter, J. R. (2009) Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ, 338, b2393-.

White, I. R. & Carlin, J. B. (2010) Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. Statistics in Medicine. In press.

John Carlin *is Director of the Clinical Epidemiology and Biostatistics Unit within the Murdoch Children's Research Institute and University of Melbourne Department of Paediatrics at the Royal Children's Hospital, Melbourne, and has a professorial appointment in the Centre for Molecular, Environmental, Genetic & Analytic Epidemiology, School of Population Health, University of Melbourne. Since completing a PhD in Statistics at Harvard University, he has had over 20 years experience working as a biostatistician. He has coauthored over 250 research publications, mainly in clinical and epidemiological journals, but also reflecting methodological research on topics in longitudinal and correlated data, and focusing recently on methods for dealing with missing data using multiple imputation. He is a founding member of the Steering Committee of the Biostatistics Collaboration of Australia, which coordinates a national postgraduate training program in biostatistics.*

MODELS FOR METHOD COMPARISON STUDIES

Bendix Carstensenr

Steno Diabetes Center, Niles Steensens Vej 2, DK-2820 Gentofter
bxc@steno.dk

Comparing two methods of measurement is normally done with limits of agreement, i.e. a prediction interval for the difference between two future measurements. In a recent publication Bland and Altman recommended the use of replicate measurements for assessment of the difference and provided examples of calculations to generalize the limits of agreement.

We will argue that the relevant approach is to set up a proper statistical model. Depending on whether replicates are exchangeable or not, different models are needed.

The point of the model we propose is that it includes the person (sample) effect as fixed, because the distribution of the sample used should not have any bearing on the prediction of future differences between methods. Many other methods proposed include the variance of the person effects. We argue that this an absurd approach.

The simplest models I propose involve standard mixed models that can be fitted with standard software. I provide a justification for the models and outline how to fit them using standard software. An extension of the model to a more realistic situation with a linear relationship between measurement methods is also introduced .

I shall demonstrate how the models can be used to provide a practical simple approach to construction of limits of agreement with replicate measures, as well as how to accommodate non-linear effects through proper transformations.

Bendix Carstensen *has been a senior statistician at Steno Diabetes Center for the last 11 years and has worked both in clinical medicine and in epidemiology.*

SAMPLING AND THE REAL WORLD

Martin Caruso

Australian Bureau of Statistics
Level 15, Exchange Plaza, Sherwood Court, Perth, WA, 6000
martin.caruso@abs.gov.au

In employer surveys the ABS uses list based frames of businesses to make inferences about variables such as employment and earnings in the targeted population of businesses. Changes to the list based frames over time and changes to how businesses are sampled can induce impacts to survey estimates that are not reflective of real world movements. In using sampling the ABS has to contend with the changes that occur over time to list based frames which necessitates sample redesigns in keeping samples optimal to deliver fit for purpose statistics. Industry classification changes, which occur about every 10-15 years, are required to keep up with new emerging industries as well as changes within industry. The ABS also employs a rotation policy to minimise respondent burden on small businesses. In 2009 a number of changes were introduced to the Average Weekly Earnings (AWE) survey, a survey which provides estimates of the earnings over employment rate. These included frame enhancements to minimise undercoverage and a new sample design to include the new ANZSIC06 industry classification, while also incorporating an improved sizing variable in stratifying the target population. To measure the impacts of these changes the old sample was run with the old sample design, old ANZSIC93 industry classification, with sample selected from the old frame. A parallel sample, maximising overlap with the old sample, was then conducted with the new sample design, new ANZSIC06 industry classification and sample selected from the new frame in the same reference period. In this I talk I will discuss the parallel sample approach that was undertaken to measure the impacts of all these changes. Also I discuss how the parallel sample is used to decompose these impacts between the old and the new sample and whether the impacts to estimates were driven by particular units or an artefact of the change processes.

References

Lawrence, D., (1991). Summary of Estimates. Australian Bureau of Statistics, Internal Report
Caruso, M.G., (2009). AWE transition from ANZSIC93 to ANZSIC06 comparison of base series, post-stratified to actual ANZSIC06 estimates. Australian Bureau of Statistics, Internal Report

Martin Caruso *currently works in the Business Survey Methodology (BSM) unit as a virtual team member based in Perth, providing methodological support to the Labour Employer Surveys (LES), Business Statistics Centre (BSC) which conduct a number of employer surveys from the Perth office. Martin has been involved in a diverse range surveys from area based household surveys, partial coverage surveys to the list based frame business surveys. His main interest is in sampling theory and its practical application to employer surveys.*

BOOTSTRAPPING SPATIAL DATA - AN APPLICATION TO ANALYSIS OF WEATHER MODIFICATION DATA

Stephen Beare[1], Ray Chambers[2], Jessie Handbury[3]

[1] Analytecon
4 Johnston Street, Narrabundah, ACT 2604
stephenbeare@mac.com

[2] Centre for Statistical and Survey Methodology
University of Wollongong, Wollongong, NSW 2522
ray@uow.edu.au

[3] Analytecon
4 Johnston Street, Narrabundah, ACT 2604
jhh2002@columbia.edu

Non-stationary spatial variation makes it extremely difficult to establish balanced real-time areas of control and effect in randomized trials of weather modification technology, and a model-based approach can be useful when evaluating the data collected in these trials. Here we describe a statistical methodology that combines regression modelling with non-parametric bootstrap simulation for this problem. This methodology has been used in an analysis of a randomized cross-over trial of a ground-based ionization (rainfall enhancement) technology known as Atlant in the Mount Lofty Ranges of South Australia in 2009. The approach is based on building a mixture-based statistical model for daily gauge-specific rainfall data that makes use of orographic and daily meteorological covariates, random gauge effects to control for unmeasured covariates and dynamically defined downwind areas to capture the level of exposure to Atlant operation. The model includes a logistic component for the probability of rain being recorded at a gauge, and a log scale linear mixed model component for the amount of rain recorded on days when there is rain. Overall, the fit of this model indicates substantial rainfall enhancement, with total rainfall downwind of the Atlant test sites over the period of the trial estimated to be somewhere between 12% to 15% higher than would have otherwise been the case. The significance of this result was confirmed by carrying out a number of non-parametric spatial bootstrap experiments that re-sampled different downwind gauges in order to build up a robust picture of the sampling distribution of the estimator of the enhancement effect.

Ray Chambers *is Professor of Statistical Methodology at the Centre for Statistical and Survey Methodology, University of Wollongong. His research interests include sample survey design and analysis, robust statistical methods and statistical modelling and inference. He has extensive experience in statistical consultation and in the application of modern statistical methods to the analysis of complex data.*

COST BENEFIT ANALYSIS OF A COMPLEX SYSTEM WITH THE DIFFERENT REPAIR/ REPLACEMENT POLICIES

Alka Chaudhary[1], Bhupendra Singh[2], Neeraj[3]

[1]Department of Statistics, Meerut College Meerut-250004, India
alka_813@yahoo.com

[2]Department of Statistics, C.C.S. University, Meerut-250005, India
bhupendra.rana@gmail.com

[3]Department of Statistics, Meerut College Meerut-250004, India

This paper deals with the analysis of a complex system with different repair/ replacement policies. The system consists of two subsystems, say A & B, connected in series. Subsystem A has two identical units with one in cold standby and B has only one unit which is non-repairable. Each unit has only two modes – normal (N) and total failure (F). The repairman is always available to repair the system and for replacement of a failure unit. Initially the system starts from one unit of subsystem A and subsystem B is operative and the other unit of subsystem A is kept as cold standby. Since the unit of subsystem B is non-repairable after failure it goes for replacement. Several measures of system effectiveness such as reliability, MTSF, steady state availability, expected profit etc., useful to system managers, are obtained by using regenerative point technique. Further, recognizing the fact that the life testing experiments are very time consuming, the parameters representing the reliability characteristics of the system/unit are assumed to be random variables. Therefore, a simulation study is also conducted for analyzing the considered system model both in classical and Bayesian set ups.

Alka Chaudhary *is an Associate Professor of Statistics at the Department of Statistics, Meerut College Meerut, Chaudhary Charan Singh University, Meerut, India. He is an active researcher and professional having expertise in modelling and analyzing engineering system models in respect of their reliability characteristics.*

MIXTURE-BASED NONPARAMETRIC DENSITY ESTIMATION WITH QUADRATIC LOSS

Chew-Seng Chee[1], *Yong Wang*[2]

[1] *The University of Auckland, Department of Statistics*
Private Bag 92019, Auckland, New Zealand
c.chee@auckland.ac.nz

[2] *The University of Auckland, Department of Statistics*
Private Bag 92019, Auckland, New Zealand
yongwang@auckland.ac.nz

Quadratic loss is predominantly used in the literature as the performance measure for nonparametric density estimation, while mixture models with a nonparametric mixing distribution have been almost exclusively studied and estimated by using the likelihood approach. In this talk, we relate both and propose an extension that estimates a nonparametric mixing distribution by minimizing the quadratic loss, and apply it, with a few variants, to nonparametric density estimation. Experimental studies show that the new estimators outperform the popular kernel-based density estimators in terms of mean integrated squared error for practically all distributions that are commonly encountered in practice, thanks to the capability of mixture models for substantial bias reduction.

Chew-Seng Chee *who originated from Malaysia is currently pursuing a Ph.D degree in Statistics at The University of Auckland under the sponsorship of the Malaysian Ministry of Higher Education and University Malaysia Terengganu. His research interests are in the areas of mixture models and model selection.*

ASYMMETRIC VOLATILITY AND DYNAMIC SKEWNESS IN BAYESIAN VALUE-AT-RISK AND EXPECTED SHORTFALL FORECASTING MODEL

Qian Chen[1], Richard Gerlach[2]

[1]Discipline of Operations Management and Econometrics, University of Sydney, Australia, Qianc@econ.usyd.edu.au

[2] Discipline of Operations Management and Econometrics, University of Sydney, Australia, R.Gerlach@econ.usyd.edu.au

A parametric method to estimate and forecast Value-at-Risk (VaR) and Expected Shortfall (ES) for a heteroskedastic financial return series is proposed. Analysis focuses on the recent popular topics of asymmetric volatility and dynamic higher moments. A smooth transition GARCH (STGARCH) models the asymmetric volatility process, in which the conditional variance complies to two different regimes with a smooth transition function continuous between zero and one. As a zero threshold is not necessary in the occurrence of regime switch, it is estimated as a parameter in this paper. To take account of potential skewness and heavy tails, the model assumes an asymmetric Laplace distribution as the conditional distribution of the financial return series. Dynamics in higher moments are captured by allowing the shape parameter in this distribution to be time-varying. The model parameters are estimated via an adaptive Markov Chain Monte Carlo (MCMC) sampling scheme, employing the Metropolis-Hastings (MH) algorithm with a mixture of Gaussian proposal distributions. The model is illustrated by a simulation study as well as an empirical experiment with four international stock market indices and two exchange rates, generating one step-ahead forecast of VaR and ES. Model comparison is investigated together with GJR-GARCH for volatility process, as well as Hansen's skewed student t distribution accounting for dynamic skewness. Standard and non-standard tests are applied to forecasts from these models. The results favor the proposed model in general.

Qian Chen *is currently a doctoral student with the Discipline of Operations Management and Econometrics, University of Sydney. Her area of interest is Bayesian Methods for Inference in financial econometric forecasting models. Qian's recent work has involved modelling the dynamics in higher moments and asymmetry in the volatility process in the heteroskedastic, heavy-tailed and skewed financial time series, with focus on the pre and post periods of the Great Financial Crisis.*

UNDERSTANDING PREDATOR-PREY RELATIONS IN FOOD WEBS USING STATISTICAL SOCIAL-NETWORK MODELS

Grace S. Chiu[1], *Anton H. Westveld*[2]

[1] *CSIRO Mathematics, Informatics and Statistics (CMIS)*
GPO Box 664, Canberra, ACT 2601, Australia
grace.chiu@csiro.au

[2] *Department of Mathematical Sciences, University of Nevada Las Vegas*
Las Vegas, NV 89154, USA
anton.westveld@unlv.edu

Recent developments in statistical methodologies for social network (SN) modelling (Gill and Swartz, 2001; Hoff et al., 2002; Westveld & Hoff, revision submitted) have helped to deepen the understanding of pairwise interactions among a network of "actors" in many scientific applications. We adapt existing methodologies to understand feeding patterns in food webs purely through empirical evidence. These methods differ from classical SN analysis techniques by explicitly modelling the random link between any given pair of species, and the complex dependence structure on these links; explanatory variables such as evolutionary distance between predator and prey can provide quantitative insight into the food web structure. The resulting statistical inference can address (i) predation activity, (ii) what makes a given species a prey or a predator, (iii) the tendency for predator-prey role reversal, (iv) trophic clustering, and (v) predation preference (latent clustering). We demonstrate our method through a meta-analysis of some famous food webs. We also explore ways to overcome modelling complications that arise due to the known, non-random absence of predation (structural zeros) between certain pairs of species, e.g. herbivores do not eat animals.

References

Gill, P.S. and Swartz, T.B. (2001). Statistical analyses for round robin interaction data. The Canadian Journal of Statistics, 29, 321–331.

Hoff, P.D., Raftery, A.E., and Handcock, M.S. (2002). Latent space approaches to social network analysis. Journal of the American Statistical Association, 97, 1090–1098.

Westveld, A.H. and Hoff, P.D. (revision submitted). A Bayesian mixed effects model for longitudinal social network data.

Grace Chiu *is currently a CMIS Senior Research Scientist. She is also an affiliate faculty member at the U of Washington (Seattle) Dept. of Statistics, and an adjunct faculty member at the U of Waterloo Dept. of Statistics and Actuarial Science. She develops statistical methodology to understand, monitor, and predict environmental phenomena, including food web structures and ecosystem health. Her work on bent-cable regression has led to various awards (for her and her Ph.D. student, respectively), various journal publications, and the R package bentcableAR. Grace is an active member of The International Environmetrics Society (TIES), an associate editor for the society journal Environmetrics, and the TIES webmaster.*

NOVEL CHROMATOGRAM ALIGNMENT METHODS FOR METABOLOMICS IN GRAPE AND WINE RESEARCH

David Clifford[1], Glenn Stone[2]

[1]CSIRO Mathematics, Informatics and Statistics
Longpocket Laboratories, Indooroopilly, QLD 4068
David.Clifford@csiro.au

[2]CSIRO Mathematics, Informatics and Statistics
Locked Bag 17, North Ryde, NSW
Glenn.Stone@csiro.au

Chromatogram alignment prior to peak identification provides a rich collection of peaks in the analysis of gas- or liquid-chromatography mass-spectrometry data (GC-MS or LC-MS). This reverses the common analysis paradigm for data of this nature where peaks are usually identified first, then clustered together based on their positions. The additional computational cost of chromatogram alignment is justified by the greater power and accuracy in peak identification when the chromatograms have been properly aligned. Moreover, all chromatogram information is retained which is necessary for biomarker discovery in untargeted metabolomics research. All forms of chromatography exhibit (sometimes appreciable) variability in retention times so solving the chromatogram alignment problem is non-trivial and of great importance to modern metabolomics research.

Aligment of GC-MS chromatograms is sometimes required prior to sample comparison and data analysis. Without alignment, direct comparison of chromatograms would lead to inaccurate results. In this talk I demonstrate a new method for computing a high quality alignment of full length GC-MS chromatograms using variable penalty dynamic time warping (Clifford et al 2009). This method aligns signals using local linear shifts without excessive warping that can alter the shape (and area) of chromatogram peaks. Software is available on CRAN for use in R (Clifford et al 2010).

References

D. Clifford, G. Stone, I. Montoliu, S. Rezzi, F.P. Martin, P. Guy, S. Bruce, and S. Kochhar. Alignment using variable penalty dynamic time warping. Analytical Chemistry, 81(3):1000–1007, 2009.
D. Clifford and G. Stone. Variable penalty dynamic time warping code for aligning GC-MS chromatograms in R. Submitted to Journal of Statistical Software.

David Clifford *is a research scientist at CSIRO Mathematics, Informatics and Statistics, and is originally from Cork, Ireland. David has a PhD in statistics from the University of Chicago for his research on the nature of spatial variation in crop yields under the guidance of Prof Peter McCullagh. David's research and work in statistics has been driven by applied and computational problems. In 2010 David received an International Science Linkages travel fellowship from the Australian Academy of Science to further develop his work on alignment methodology through a collaborative visit with Dr Ron Wehrens at Istituto Agrario San Michele all'Adige in Northern Italy.*

PREDICTING TRIHALOMETHANE FORMATION IN DRINKING WATER USING RANDOM FORESTS

R. Gupta[1], R. Collinson[1], A. Heitz[2], R. Trolio[3] and A. Lethorn[3]

[1]Department of Mathematics and Statistics, Curtin University of Technology
Kent Street Bentley, WA 6845, Australia
r.guta@curtin.edu.au, r.collinson@curtin.edu.au

[2]Curtin Water Quality Research Centre, Department of Applied Chemistry
Curtin University of Technology, Kent Street Bentley, WA 6845, Australia
a.heitz@curtin.edu.au

[3]Water Corporation of Western Australia
629 Newcastle Street Leederville, WA, 6007, Australia
Rino.Trolio@watercorporation.com.au, Arron.Lethorn@watercorporation.com.au

Although disinfection of water is essential for producing healthy potable water, the disinfectants react with natural organic matter to form undesirable by-products such as trihalomethanes (THMs). THMs can be potentially harmful to human health and it is desirable that their concentrations be minimised. In Australia, total THM (TTHM) concentrations in drinking water are subject to guideline values, with a guideline maximum TTHM concentration, based on health considerations, of 250 µg L-1. The current methods of monitoring these by-products used by the Water Corporation of Western Australia (WCWA) take a minimum of nine days. This is too time consuming and inadequate for timely management decisions. Thus an accurate and time-efficient technique needs to be developed. In this paper, we describe a statistical model for prediction of THM concentrations using rapidly obtainable water quality data. A range of statistical modelling methods were investigated for estimating the total THMs in drinking water. The best of these uses random forests which is an ensemble classification and regression method. Each tree is constructed using a different bootstrap sample of the data based on a selection of small number of the input variables and final prediction is obtained by aggregating over the ensemble. Prediction accuracy is improved by using ensembles of trees. The final proposed model is cost efficient, has excellent prediction accuracy and has the potential to produce same day results. The model is currently being used by WCWA to monitor water quality across Western Australia. The predictions from statistical models till date have been within the acceptable range.

References

Breiman, L. (2001), 'Random Forests', Machine Learning, vol. 45, no. 1, pp. 5-32.
Gupta, R., Collinson, R., Heitz, A., Warton, B., Trolio, R., Maus, A. and McNeil, S. (2009), 'Predicting Trihalomethane formation in drinking water using statistical models', in OzWater2009.
Obolensky, A. and Singer, P.C. (2008), 'Development and Interpretation of Disinfection Byproduct Formation Models Using the Information Collection Rule Database', Environmental Science & Technology, vol. 42, no. 15, pp. 5654-5660.

Roger Collinson *is currently Research Fellow in the Department of Mathematics and Statistics at Curtin University of Technology. He lectures in mathematics and statistics and also works for the Statistical Consulting Unit (SCU) within the department. He has been involved in a range of projects including statistical modelling of water quality and reserves estimation in petroleum engineering. He obtained BSc(Hons) in 1993 and PhD in applied mathematics in 1998, both from Curtin University of Technology. Prior to his current appointment, he worked in the Western Australian Centre of Excellence in Industrial Optimisation on effective optimization techniques for open-pit mine scheduling problems.*

MODELLING GEOGRAPHIC DISPARITIES IN CANCER OUTCOMES IN QUEENSLAND

Susanna Cramb[1], Kerrie Mengersen[2], Peter Baade[3]

[1] Viertel Centre for Research in Cancer Control
Cancer Council Queensland and Queensland University of Technology
553 Gregory Tce, Fortitude Valley Queensland 4006
susannacramb@cancerqld.org.au

[2] Discipline of Mathematical Sciences, Queensland University of Technology
George St, Brisbane Queensland 4000
k.mengersen@qut.edu.au

[3] Viertel Centre for Research in Cancer Control, Cancer Council
553 Gregory Tce, Fortitude Valley Queensland 4006
peterbaade@cancerqld.org.au

Cancer is a term used to describe a variety of malignant neoplastic diseases, each with separate aetiology and prognosis. Strong evidence exists for discrepancies in cancer outcomes by location, both internationally and within Australia. Although Queensland is the most decentralized state in Australia, detailed investigations of the spatial distribution of cancer incidence and survival have not previously been conducted. To effectively address potential inequalities they must first be quantified at the small area level, so Statistical Local Areas (SLAs) were used as the area-level boundaries (n=478). When examining these small areas, data sparseness and potentially dependent spatial patterns were important considerations. Therefore, Bayesian hierarchical models were used to produce reliable and robust estimates of incidence and survival for the leading 20 cancers by gender over the time period 1998-2007. Incidence was modelled using the Besag, York and Mollié model (Besag et al. 1991) while survival was modelled using a relative survival (excess mortality) model similar to that proposed by Dickman et al (2004). Methodological issues in modelling these data, as well as the results (which were published as the Cancer Atlas of Queensland) will be discussed. In addition, classification and regression tree analyses were conducted to examine which area-wide characteristics are potentially influencing disparities in cancer diagnosis. Results from these analyses and future research directions will be presented.

References

Besag, J., York, J. & Mollie, A. (1991) Bayesian image restoration, with two applications in spatial statistics. Ann. Inst. Statist. Math., 43, 1-59.

Dickman, P. W., Sloggett, A., Hills, M. & Hakulinen, T. (2004) Regression models for relative survival. Stat Med, 23, 51-64.

Susanna Cramb *is employed as a Research Officer at the Viertel Centre for Research in Cancer Control, Cancer Council Queensland, as well as studying towards a PhD through the Discipline of Mathematical Sciences, Queensland University of Technology. In addition to being lead author on the recently released Cancer Atlas of Queensland, Susanna has co-authored several comprehensive reports on individual cancers (lung, colorectal and breast cancers). Her research interests include disease mapping, cancer disparities and Bayesian methodology.*

MIXTURE OF RANDOM EFFECTS FOR INDIVIDUAL LEARNING CURVES

Edward Cripps[1], *Robert Wood*[2], *Sally Wood*[3]

[1] *School of Mathematics, University of Western Australia, WA, 6009*
ecripps@maths.uwa.edu.au

[2] *Melbourne Business School, University of Melbourne, Victoria 3053*
r.wood@mbs.edu

[3]*Melbourne Business School, University of Melbourne, Victoria 3053*
sally.wood@mbs.edu

This presentation applies a mixture of random effects model to a current topic in the psychological literature where it has been proposed that individuals belong to one of two groups: entity theorists who believe ability is innate and incremental theorists who believe ability is an acquired set of skills. Psychological theory argues that entity theorists are more prone to negative emotions following setbacks when learning new tasks, which undermines their information processing and leads to downward spirals in performance. To investigate this hypothesis the Accelerated Learning Laboratory, Melbourne Business School, has conducted experiments in which individuals from both groups were subjected to repeated tasks and their performances evaluated. The data is longitudinal and to incorporate spiralling behaviour we extend the standard random effects model to a mixture of possibly two random effects. Inference is Bayesian and computations are carried out using a Markov chain Monte Carlo algorithm. In effect we have two models, the first model is that a spiral exists and the second model is that it does not. If the individual spirals we average over another class of models, which are the possible locations of the spiral. So we are averaging over two levels: the first level being spiral or no spiral, and the second level being the location of the spiral. We weight all these possible models by their posterior probability. The results support the hypothesis that entity theorists are more prone to spiralling behaviour.

Edward Cripps *is Assistant Professor at the School of Mathematics and Statistics, University of Western Australia. His research interests include computational statistics, Bayesian methods and applications, longitudinal data, mixture modelling, model averaging and environmental statistics.*

PRODUCING STATE LEVEL ESTIMATES OF HOURS WORKED

Philip Crouch

Australian Bureau of Statistics
philip.crouch@abs.gov.au

The Australian Bureau of Statistics (ABS) publishes estimates of total hours worked by employed Australians, which complement the employment rate as a measure of economic activity and employment in Australia. They are also used as a labour input for the productivity estimates in the Australian System of National Accounts. Recently, there has been increased interest in using these data to measure the nature of, and the recovery from, the global financial crisis and its effect on the Australian labour force. In meeting user demands for finer level estimates, the ABS has improved methods of accounting for moving holiday effects in these time series and has developed aggregation techniques in higher dimensions.

Currently, total hours worked by employed Australians are estimated at the Australian level, decomposed into a two dimensional aggregation structure where the marginal totals of sex and employment status are forced to sum to the Australia total by two dimensional reconciliation. Finer estimates have been produced, using state as an additional variable. Producing finer level estimates requires different methods in accounting for moving holidays and ensuring additivity in seasonally adjusted estimates.

Holiday correction, in this context, is a process of removing the effects of regular non-random events, and is performed prior to adjusting for monthly seasonality. The holiday corrected series will therefore be without known holiday impacts, and considered to reflect hours worked for a 'standard working week'. Holiday corrected stock series of hours worked, derived from the ABS Monthly Population Survey (MPS), are used to produce monthly flow estimates of 'standard working weeks' using a linear interpolation based methodology. Holiday effects are then reintroduced into these estimates to give monthly flow estimates of actual hours worked.

Methods are given for producing flow estimates that reflect actual hours worked.

Philip Crouch *is an early career statistician working in the Time Series Analysis section of the Australian Bureau of Statistics.*

RISK PREDICTION MODELS IN CARDIOVASCULAR DISEASE RESEARCH

Jisheng Cui

Department of Health
50 Lonsdale Street, Melbourne, VIC 3000, Australia
jisheng.cui@health.vic.gov.au

Many risk prediction models have been developed for cardiovascular diseases in different countries during the past three decades. However, there has not been consistent agreement regarding how to appropriately assess a risk prediction model, especially when new markers are added to an established risk prediction model. Researchers often use the area under the receiver operating characteristic curve (ROC) to assess the discriminatory ability of a risk prediction model. However, recent studies suggest that this method has serious limitations and cannot be the sole approach to evaluate the usefulness of a new marker in clinical and epidemiological studies. To overcome the shortcomings of this traditional method, new assessment methods have been proposed. The aim of this article is to overview various risk prediction models for cardiovascular diseases, to describe the receiver operating characteristic curve method and discuss some new assessment methods proposed recently.

Dr Jisheng Cui *is a Senior Biostatistician at the Department of Health, Melbourne. He has published over 70 peer-reviewed journal articles and is interested in statistical methods for the analysis of longitudinal data, survival data and recurrent event data. Dr Cui has been awarded several large NHMRC and NIH funded research projects, either as a first chief investigator or as a collaborative chief investigator. Dr Cui has been invited to give over 30 seminars and conference presentations, especially by Fred Hutchinson Cancer Research Center in Seattle, MD Anderson Cancer Center in Houston as well as the University of Alberta in Canada. Dr. Cui is a statistical reviewer for the Medical Journal of Australia.*

QUANTIFYING THE ASSOCIATION BETWEEN CLIMATE AND BIOLOGICAL INDICATORS USING WAVELET ANALYSIS

Ross Darnell[1], *Melissa Dobbie*[2]

[1] *CSIRO Mathematics, Informatics and Statistics*
CSIRO Marine Laboratories, P O Box 120, Cleveland QLD 4163
Ross.Darnell@csiro.au

[2] *CSIRO Mathematics, Informatics and Statistics*
120 Meiers Road, Indooroopilly QLD 4068
Melissa.Dobbie@csiro.au

We will demonstrate the advantages of wavelet analysis to quantify the relationships between ecological and environmental time series. Environmental time series measurements are often recorded at different frequencies and exhibit non-stationary properties. By partitioning the time series into physically meaningful components using wavelet analyses (Percival et al, 2010), relationships between responses and forcing variables that are recorded at different temporal scales can be quantified.

References

Percival, D.B., Lennox, S.M., Wang, Y.-G. and Darnell, R.E. (2010). Wavelet-based multiresolutional analysis of Wivenhoe Dam water temperature. Water Resources Research.(submitted).

Ross Darnell *is a team leader with the CSIRO Mathematics, Informatics and Statistics, focusing on statistical modelling of aquatic ecosystems.*

ANALYSIS OF PROBABILISTICALLY LINKED SURVEY AND ADMINISTRATIVE DATA; AN EMPIRICAL APPLICATION TO THE 2006/7 NEW ZEALAND HEALTH SURVEY

Walter R. Davis

Statistics New Zealand
Statistics House, Harbour Quays, Wellington, New Zealand
walter.davis@stats.govt.nz

A greater use of administrative data is a continuing trend in official statistics. A particular area of interest is the linking of survey data to administrative data to expand the potential for analysis and to reduce respondent burden. However methodology to analyse such data while controlling for both the complex sample design and probabilistic linkage and non-linkage errors has been a hindrance. This presentation applies techniques for the estimation of linear models for probabilistically linked survey and administrative data developed by Chambers (2009) and Kim and Chambers (2010) to a provisional dataset formed by linking the 2006/7 New Zealand Health Survey with healthcare system usage administrative data (e.g. hospitilisation). This presentation focuses on applying the technique in a real-world setting, briefly describes a SAS macro written for estimation and presents an extension of the model for longitudinal fixed effects analysis.

References

Chambers, R 2009. "Regression Analysis of Probability-Linked Data," Official Statistics Research Series, 4. Available from www.statisphere.govt.nz/official-statistics-research/series/default.htm.
Kim, G and Chambers, R 2010. "Regression Analysis using Longitudinally Linked Data." Presented at the Australian Statistical Conference 2010, Fremantle.

Walter R. Davis *is a Principal Methodologist at Statistics New Zealand. His areas of interest include survey statistics, official statistics, measurement models and longitudinal analysis. His current research includes work on the analysis of linked data, survey non-response and measurement properties of health and other wellbeing scales.*

EXPLORING METEOROLOGICAL CONDITIONS RELATED TO WIND FARM POWER VARIABILITY USING MACHINE LEARNING

Robert J. Davy[1], Milton J. Woods[2], Christopher J. Russell[3], Alberto Troccoli[4], Peter A. Coppin[5]

[1] CSIRO Marine and Atmospheric Research, GPO Box 3023, Canberra ACT 2601, Australia
robert.davy@csiro.au

[2] CSIRO Marine and Atmospheric Research, GPO Box 3023, Canberra ACT 2601, Australia
milton.woods@csiro.au

[3] CSIRO Marine and Atmospheric Research, GPO Box 3023, Canberra ACT 2601, Australia
chris.russell@csiro.au

[4] CSIRO Marine and Atmospheric Research, GPO Box 3023, Canberra ACT 2601, Australia
alberto.troccoli@csiro.au

[5] CSIRO Marine and Atmospheric Research, GPO Box 3023, Canberra ACT 2601, Australia
peter.coppin@csiro.au

Measurements have shown that there are certain time periods when the wind is highly variable across southern Australia. Under such conditions the risk of a large rapid change in the output of wind farms is increased. This has the potential to impact upon electricity supply quality and reliability. Wind variability tends to increase with mean wind speed, however a large component of this is not explained by the wind speed. In this work, gridded spatial meteorological fields are statistically downscaled to model wind power variability at coastal locations in southern Australia. These fields are each decomposed using principal components analysis (PCA) prior to the empirical modelling process. Data representing whole years are used to form a model which is then validated using independent data from another year. The principal components which best predict wind power variability are examined, providing insight into the phenomenon of wind power variability in this region. To allow for nonlinearity and complex interaction between variables, all empirical models are built using machine learning techniques.

Robert Davy *currently works as a data analyst at the Weather and Energy Research Unit at CSIRO Marine and Atmospheric Research, Canberra. He is interested in aspects of forecasting for renewable energy. He has been involved in consultation for both the Australian Government and the wind energy industry.*

ISSUES IN MIXED MODEL ANALYSES OF LONGITUDINAL LUNG FUNCTION DATA

Nicholas de Klerk

Centre for Child Health Research, University of Western Australia
100 Roberts Road, Subiaco, WA 6008, Australia
nickdk@ichr.uwa.edu.au

One of the best ways of assessing possible health effects of different occupations is by monitoring new employees as they progress through a particular industry. While carcinogenic effects of potential hazards can only be expected to occur long after exposure, other health effects may occur sooner. Because of their exposure, albeit low, to known harmful dusts, miners and refinery workers need to be monitored for their respiratory health. Because such monitoring involves regular repeated measures on the same people, analyses using generalized mixed models are an obvious choice to examine health effects of exposures. These studies are also complicated by decisions as to how and when to measure which potentially harmful exposures. Decisions also have to be made about how to measure non-occupational exposures (such as smoking) and which other confounders to try to adjust for in analyses and how to model them. This study describes the different and potentially conflicting results that arose from analyses of such an inception cohort study in the aluminium industry, where effects on both categorical (symptoms) and continuous (measures of lung function) responses were examined for several exposures. In analyses with continuous outcomes, the different choices as to how to model age in both random and fixed parts of these models (including using fractional polynomials) were crucial in terms of decision making, but were difficult to distinguish on purely statistical grounds. As is often the case in such observational data, the most reliable results only arose after numerous 'feedback loops' between clinicians, epidemiologists, hygienists and biostatisticians.

References

Cui J, de Klerk N, Abramson M, et al. Fractional polynomials and model selection in generalized estimating equation analysis, with an application to a longitudinal epidemiological study in Australia. Amer J Epidemiol 2009;169:113-121.

Musk AW, de Klerk NH, Beach JR, et al. Respiratory symptoms and lung function in alumina refinery employees. Occup Env Med 2000;57:279-283

Nick de Klerk *currently leads the Biostatistics and Bioinformatics Department at the telethon Institute fro Child Health Research and is an Adjunct Professor in the School of Population Health, University of Western Australia. His main areas of interest are in epidemiological methods particularly in child health, occupational health, study design, clinical trials, biostatistics, and the analysis of microarray and other genetic data.*

STATISTICAL METHODOLOGIES IN THE PHARMACEUTICAL INDUSTRY: APPLICATIONS TO DRUG REIMBURSEMENT

Philippa Delahoy

Pfizer Australia Pty Ltd
38-42 Wharf Road, West Ryde, NSW 2114
philippa.delahoy@pfizer.com

In Australia, as well as other countries such as the UK and Canada, pharmaceutical companies prepare and submit reimbursement applications in order to list products on the national formulary (Pharmaceutical Benefits Scheme). Reimbursement submissions are based on the principles of evidence-based medicine, requiring inclusion and review of the entire evidence base of each product, presented in the context of a 'comparator product' defined as the product most likely to be replaced in clinical practice (active drug or placebo/best supportive care). Information is presented on the comparative effectiveness, tolerability and utility of the new product under consideration, in order to derive an estimate of cost-effectiveness. The analysis techniques required to form each of these components may include meta-analysis, meta-regression, indirect comparisons and data extrapolation; all of which aim to translate results from the clinical trials setting to the Australian environment. The role of the statistician is not only to conduct these analyses, but to provide critical appraisal of the numerous data sources, to advise on the robustness of the analyses, to perform sensitivity analyses and to supply inputs into the economic model through which cost-effectiveness is derived. With no two reimbursement submissions being the same, the work of the statistician in this area is challenging as well as rewarding and interesting.

Philippa Delahoy *currently works as an Outcomes Research Statistical Lead with Pfizer Australia, in Sydney. Her area of interest is in estimating the cost-effectiveness of innovative medicines for the purposes of reimbursement, with a particular interest in the methodologies employed for reimbursement of oncology products. Philippa is the current chair of the Australian Pharmaceutical Biostatistics Group; a not-for-profit association of pharmaceutical industry statisticians, whose mission is to ensure high statistical standards to assist in the decision processes which provide safe, efficacious and cost-effective pharmaceutical products for the health and quality of life of people in Australia.*

UNDERSTANDING THE GENETIC COMPONENTS OF NIR SPECTRA

Diepeveen D[1], Clarke G.P.Y[2], Bellgard M[3], Appels R[4]

[1] Department of Agriculture and Food, Government of Western Australia
Locked Bag 4, Bentley Delivery Centre, Bentley 6983, Western Australia
dean.diepeveen@agric.wa.gov.au

[2] Department of Agriculture and Food, Government of Western Australia
Locked Bag 4, Bentley Delivery Centre, Bentley 6983, Western Australia
peter.clarke@agric.wa.gov.au

[3] Centre for Comparative Genomics, Murdoch University
Murdoch University, Murdoch, 6150, Western Australia
mbellgard@ccg.murdoch.edu.au

[4] Centre for Comparative Genomics, Murdoch University
Murdoch University, Murdoch, 6150, Western Australia
rappels@ccg.murdoch.edu.au

Near Infrared Spectroscopy (NIR) is commonly used in plant breeding to screen grain quality traits. The predictions are based on calibration curves developed using NIR spectrum features which are correlated with a grain trait. Common traits include protein, grain-size, grain-weight and grain-colour. Recent research suggests that to better differentiate between plant breeding germplasm, an understanding of the relationships between heritable spectral positions within a spectrum and the grain-trait is required. A narrow genotype doubled haploid population and a diverse genotype x environment datasets are used to estimate NIR absorbance predictions and heritability for each spectral point of a spectrum. Subsequent multivariate analyses using the environment-removed NIR absorbance predictions provide greater discrimination of plant breeding germplasm for commonly measured grain quality traits.

Dean Diepeveen *is currently a researcher with the cereal grain quality group at DAFWA. His interests are in data integration, analyses and screening of plant breeding germplasm. He has worked in several research areas with the plant breeding supply chain at DAFWA over the last 16 years. Prior to joining DAFWA, he worked for 10 year as researcher at UWA in the mathematics and medical statistics area. During the last 4 years he has undertaken PhD studies on implementing a data integration framework in wheat breeding.*

BOOTSTRAP RESAMPLING AND APPLICATIONS TO HIGH-DIMENSIONAL BIOINFORMATICS DATA IN CANCER RESEARCH

Bradley M. Broom[1], Kim-Anh Do[2]

[1] Department of Bioinformatics and Computational Biology
U.T.M.D. Anderson Cancer Center, 1515 Holcombe Blvd, Houston, TX 77401, USA
bmbroom@mdanderson.org

[2] Department of Biostatistics
U.T.M.D. Anderson Cancer Center, 1515 Holcombe Blvd, Houston, TX 77401, USA
kimdo@mdanderson.org

Many problems in bioinformatics involve several thousands of variables, but only have a few hundred data points, if that. Bootstrapping techniques are a vital strategy for generating high quality results and eliminating spurious conclusions. In this talk we briefly describe some novel applications of bootstrapping to bioinformatics algorithms and discuss the consequent improvement in the results obtained. First, we apply bootstrap aggregation to gene shaving to learn robust clusters of co-expressed genes. On a glioma dataset, bagging improved the R^2 on a separate test set from 38% to 59%. Second, we use bootstrap aggregation to learn robust Bayesian network models from gene expression breast cancer data. We show that the bootstrap aggregation of complex model learners obtains results equivalent to fully Bayesian methods of network inference. Finally, we describe how bootstrap resampling can be used in gene correlation experiments to identify outliers and obtain better estimates of the true correlations.

Kim-Anh Do, *Ph.D is a Professor in Biostatistics who has made significant scientific contributions by developing and evaluating a number of novel computer-intensive statistical method, including methods for improving the computational efficiency of the bootstrap, empirical likelihood and Bayesian mixture modelling of gene expression and proteomic profiles, with significant applications to biomedical and cancer research. She has received a Faculty Scholar Award at M.D. Anderson Cancer Center and has been a recipient of large and small grants from the Australian Research Council, the Australian Academy of Science, the National Cancer Institute (USA), and the National Institute of Health (USA). She is an elected Fellow of the American Statistical Society and a Fellow of the Royal Statistical Society.*

ON VOLATILITY ESTIMATION FOR THE HIGH FREQUENCY STOCK PRICES

Nikolai Dokuchaev

Department of Mathematics and Statistics, Curtin University of Technology
GPO Box U1987, Perth, Western Australia, 6845
N.Dokuchaev@curtin.edu.au

We consider a problem of volatility estimation for high frequency time series of prices generated by the continuous time stochastic Ito equations with time variable parameters that are not directly observable. The problem is extremely important for pricing of derivatives on optimal portfolio selection. This is why this problem was intensively studied; there are many well developed algorithms. We suggest some technical modifications that may help to reduce the estimate error. In particular, we present some special linear transformations with causal integral kernels that preserve the volatility and reduce the impact of the presence of time variable and unknown appreciation rates. In addition, we suggest a modification of the standard formulae for volatility that reduces the error by using some features of the Ito process generating the high frequency data.

Nikolai Dokuchaev *currently works as associate Professor in the Department of Mathematics and Statistics, Curtin University, in Perth. His area of interest is Stochastic Analysis, Statistical and Mathematical Finance, Stochastic Control, and Signal Processes. Nikolai published a number of papers and two monographs in these areas.*

THE ANALYSIS OF TIME-USE DATA: AN ALTERNATIVE TO TOBIT MODELS

Peter K Dunn[1], *Jude Brown*[2]

[1] *Faculty of Science Health and Education, University of the Sunshine Coast*
Maroochydore DC Queensland 4558
pdunn2@usc.edu.au

[2] *School of Cognitive, Behavioural and Social Sciences, University of New England*
Armidale NSW 2351
jude.brown@une.edu.au

Time-use data (TUD) records the time spent on everyday activities. TUD is both discrete (with exact zeros when the activity is not done) and continuous, and so Tobit models are almost always used for analysing TUD. In this talk, we consider TUD from the first wave of the Longitudinal Study of Australian Children, which records the activities of infants and four-year-old children. We present Tweedie generalized linear models as an alternative to Tobit regression models, and compare the two types of models. Theoretical advantages of the Tweedie models include: a sensible interpretation; the flexibility of three parameters; the exact zeros are modelled explicitly. We then compare the two models from a practical viewpoint by examining the time spent by four-year-olds watching television, walking, and driving in the car. The results show that the Tweedie model fits the data at least as well as the Tobit models, and so are a viable alternative to Tobit models for the analysis of TUD.

References

Dunn, P.K., and Smyth, G.K. (2008). Evaluation of Tweedie exponential dispersion model densities by Fourier inversion. Statistics and Computing, 18, 73-86.

Brown, J., and Dunn, Peter K. (2006). Analysis of LSAC time use data: Opportunities and challenges.Paper presented at the ACSPRI Social Science Methodology Conference, University of Sydney, 10-13 December.

Dunn, P.K., and Smyth, G.K. (2004). Series evaluation of Tweedie exponential dispersion model densities. Statistics & Computing, 15(4), 267-280.

Dr Peter Dunn *is a Biostatistician in the Faculty of Science, Health and Education at the University of the Sunshine Coast, Queensland.*

SOME LESSONS ABOUT THE ODDS RATIO: OLD AND NEW

Sue Finch[1], *Ian Gordon*[2]

[1] *Statistical Consulting Centre, The University of Melbourne*
Parkville, Australia, 3010
sfinch@unimelb.edu.au

[2] *Statistical Consulting Centre, The University of Melbourne*
Parkville, Australia, 3010
irg@unimelb.edu.au

Many people find the notion of "odds" difficult to understand, although it is common in betting contexts. The "odds ratio"—a ratio of ratios—can be even more puzzling (e.g. Deeks, 1998). We demonstrate misunderstandings of the meaning of the odds ratio, and present evidence that misinterpretations can appear in high quality journals. We point out a connection between the odds ratio and a more straightforward measure, which offers useful insight, and can be regarded as a new characterization of the odds ratio; we are not aware of it being documented previously. We illustrate the use and interpretation of this connection, and propose that researchers consider this connection in trying to make substantive sense of their findings.

References

Deeks JJ. (1998). When can odds ratios mislead? British Medical Journal. 317:1155-1156.

Sue Finch *is a consultant at the Statistical Consulting Centre at The University of Melbourne. She works with researchers across a broad range of disciplines. Her interests include the use of multimedia in teaching statistics, and finding better ways of communicating and understanding statistical concepts.*

EFFICIENT BAYESIAN ESTIMATION OF THE MULTIVARIATE DOUBLE CHAIN MARKOV MODEL

Matthew Fitzpatrick[1], *Dobrin Marchev*[2]

[1] *School of Mathematics and Statistics, The University of Sydney, NSW 2006, Australia*
mfit4623@uni.sydney.edu.au

[2] *School of Mathematics and Statistics, The University of Sydney, NSW 2006, Australia*
dobrin@maths.usyd.edu.au

The double chain Markov model (DCMM) is used to model observable data $\{Y_t\}_{t=1}^n$ as a Markov chain with transition matrix P_{x_t} dependent on the value of an unobservable (hidden) Markov chain $\{X_t\}_{t=1}^n$. We extend the method presented in Chib (1996) for sampling from the posterior distribution associated with the DCMM when the observed process $\{Y_t\}_{t=1}^n$ consists of vectors of (possibly) different lengths.

Convergence of the Gibbs sampler, used to simulate the posterior density, is improved by adding a random permutation step as described in Frühwirth-Schnatter (2001). The post processing algorithm of Stephens (2000) is used to overcome any identifiability issues that arise from label-switching. Simulation studies and an application to real data representing the credit rating dynamics of financial companies are presented. We show that when fitting our model to the observed credit rating transitions of a large portfolio of financial companies, the hidden regimes selected over time bear remarkable resemblance to many of the economic events in the last 30 years.

References

Chib, S. (1996). Calculating posterior distributions and modal estimates in Markov mixture models. Journal of Econometrics, 75, 79-97.

Frühwirth-Schnatter, S. (2001). Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. Journal of the American Statistical Association, 96, 194-209.

Stephens, M. (2000). Dealing with label switching in mixture models. Journal of the Royal Statistical Society. Series B. Statistical Methodology, 62, 795-809.

Matthew Fitzpatrick *is a post-graduate research student at the School of Mathematics and Statistics at the University of Sydney. He is interested in researching the theory and applications of Markov mixture models, with a particular interest in their applications to credit rating migration models. Matthew also works in the Banking and Finance industry in the Quantitative Risk Tools and Modelling department of the Commonwealth Bank, where he is involved in the development and application of various statistical techniques for credit risk measurement.*

MODELLING AND ESTIMATION FOR BIVARIATE FINANCIAL RETURNS

Thomas Fung[1] *and Eugene Seneta*[2]

[1] *Department of Statistics, Faculty of Science, Macquarie University, N.S.W. 2109, Australia*
thomas.fung@mq.edu.au

[2] *School of Mathematics and Statistics, University of Sydney, N.S.W. 2006, Australia*
eseneta@maths.usyd.edu.au

Maximum likelihood estimates are obtained for long data sets of bivariate financial returns using mixing representation of the bivariate (skew) Variance Gamma and (skew) t distributions. By analysing simulated and real data, issues such as asymptotic lower tail dependence and competitiveness of the two models are illustrated. A brief review of the properties of the models is also included. The present paper is a companion to papers by Demarta and McNeil (2005) and Finlay and Seneta (2008).

References

Demarta, S. and McNeil, A.J. (2005). The t copula and related copulas. International Statistical Review, 73, 111-129. Analysis of Longitudinal Data. Oxford: Clarendon Press.
Finlay, R. and Seneta, E. (2008). Stationary-increment Variance-Gamma and t models: Simulation and parameter estimation. International Statistical Review, 76, 167-186.
Fung, T. and Seneta, E. (2010). Modelling and Estimation for Bivariate Financial Returns. International Statistical Review, 78, 117-133.

Thomas Fung *is a recent Ph.D. graduate from the University of Sydney and currently a Lecturer in the Department of Statistics at Macquarie University in Sydney. His area of interest is to study Variance Gamma (VG) distributions and their applications in comparison with other heavy tail distributions in the area of financial modelling. Currently, his research is focused on multivariate analysis and distribution theory in financial related areas.*

CLASSIFICATION OF IMBALANCED DATA SETS BY USING OVER-SAMPLING ALGORITHMS BASED ON COMPLEXITY MEASURE

Siva Ganesh, Nafees Anwar, Geoff Jones and Selvanayagam Ganesalingam

Massey University Inst. of Fundamental Sciences
Private Bag 11222, Palmerston North 4442, New Zealand
s.ganesh@massey.ac.nz

There are a number of aspects that might influence the performance achieved by existing classifiers. It has been reported in the literature that one of these aspects is 'Class Imbalance' in which one class has many more instances than others. Our experiments provide evidence that class imbalance does not necessarily deter the performance of a classifier, rather it is related to too few minority class examples in the presence of other complicating factors, such as class overlapping.

This paper presents a new learning approach for classification applications involving imbalanced data sets using our proposed complexity measure. We use a nearest neighbour approach for characterising the complexity of classification problems. We have studied the comparative advantages of two methods, Euclidean distance and proximity measure by Random Forest, for constructing nearest neighbours. We have investigated a collection of two-class problems from the UCI repository, and observed that there are strong correlations between classifier accuracy and class overlap. The experiments have also demonstrated that the similarity measure by Random Forest is compatible with Euclidean distance for numerical data but has obvious advantages for data consisting of categorical or mixed type of variables where Euclidean distances cannot be computed directly.

In our approach, a clustering technique is employed to resample the original training set based on the proposed complexity measure of the minority class training examples. Based on the complexity measure we over-sample minority class examples by using active over sampling techniques like SMOTE (2002) or over sampling through PCA (principal component analysis) and produce sub-classes with relatively balanced sizes. This approach has been implemented and tested on benchmark (UCI) data sets. Experimental results show that with the proposed learning approach, it is possible to tackle the class imbalance problem, without compromising the overall classification performance.

References

http://archive.ics.uci.edu/ml/datasets.html
Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2002), SMOTE: Synthetic Minority Over-sampling Technique, Journal of Artificial Intelligence Research, 16, 321–357.

Siva Ganesh *is a Senior Academic in Statistics at the Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand. He has specific expertise in Applied Statistics, Data Mining and Statistical Computing and Consulting. He has been involved in university teaching including running workshops and short courses in New Zealand and internationally, in the areas of multivariate statistics, experimental designs and analysis, biostatistics and data mining. His current research interests include, absolute-value discriminant analysis, classification problems involving class-imbalance, feature selection, clustering and visualization, and predicting RNA-infrastructure via visualizing integrated regulated protein networks.*

DEFINING COMMON RESPIRATORY AND ALLERGIC MANIFESTATIONS OVER TIME: A LATENT CLASS ANALYSIS

Frances Garden[1], Judy Simpson[2], Catarina Almqvist[3], Euan Tovey[4], Guy B Marks[5]

[1] Sydney School of Public Health, The University of Sydney/Woolcock Institute of Medical Research
Room 128C, Edward Ford Building A27, The University of Sydney, NSW, 2006
frances.garden@sydney.edu.au

[2] Sydney School of Public Health, The University of Sydney
Edward Ford Building A27, The University of Sydney, NSW, 2006
judy.simpson@sydney.edu.au

[3] Department of Medical Epidemiology and Biostatistics, Karolinska Institutet
SE-171 77 Stockholm, Sweden
catarina.almqvist@ki.se

[4] Woolcock Institute of Medical Research
431 Glebe Point Road, Glebe, NSW 2037
ert@med.usyd.edu.au

[5] Woolcock Institute of Medical Research
431 Glebe Point Road, Glebe, NSW 2037
guy.marks@sydney.edu.au

The relationship between early life exposures and the development of asthma and allergic disease in children is complex. Our aim was to define classes of common respiratory and allergic manifestations measured over the first 8 years of life and determine risk factors for these classes. We used data from the Childhood Asthma Prevention Study – a birth cohort at high risk for asthma. Data on symptoms of allergic disease (wheeze, cough, rhinitis, eczema), diagnosed illnesses, skin prick tests (SPT) to common allergens, blood tests and lung function were collected at 1.5, 3, 5 and 8 years. We used latent class models (latent trajectory and latent markov models) to define the classes. Using latent class models allows us to take into account the patterns of change over time and utilise both the cross-sectional and longitudinal nature of the data. The latent class trajectory analysis revealed that a three class model fitted the SPT data best. Subjects were classed as non-atopic, atopic to house dust mite only or atopic to multiple allergens. Including subject random effects in the SPT latent class trajectory models further refined the classes; for example, in the 3 class model the groups could be described as “non-atopic”, “house-dust mite and grass allergies” and “peanuts and multiple allergies”. This talk will describe the resultant class structures from extending the SPT model to include period random effects, and other symptoms and illnesses of asthma and allergic disease. We will also discuss how well these classes can be predicted by risk factors such as sex, tobacco smoke exposure and genetic information.

Frances Garden *is a PhD student with the Sydney School of Public Health and the Woolcock Institute of Medical Research. The aim of her PhD is to apply appropriate sophisticated statistical models to explore the complex relationship between early life exposures and the development of asthma and allergic disease in children.*

QUANTILES VERSUS DENSITIES: A CASE STUDY OF SIMILARITIES AND DIFFERENCES

Gibbons K[1], MacGillivray H[2]

[1] Mater Medical Research Institute
Level 2, Quarters Building, Annerley Road, Woolloongabba Q 4102
kristen.gibbons@mater.org.au

[2] Mathematical Sciences, Queensland University of Technology
Gardens Point Campus, GPO Box 2434, Brisbane Q 4001
h.macgillvray@qut.edu.au

New distributional families have been, and continue to be, sought for both modelling and fitting data across theoretical and applied statistics. Some of these families are generated through a probability density approach, and more recently, exploratory data analysis has contributed to distributional families arising from transformations of a base distribution. This work considers three families; one generated through a density function approach from the symmetric-*t* (the skew-*t* distribution), and the other two generated through extensions and generalizations of the Tukey style of transformation approach of the standard normal (the generalized *g*-and-*h* and the *g*-and-*k*).

The motivation of the work is to explore, apply and compare the three families in data analysis, including fits to univariate data sets and residuals that are nonnormal and are of interest in a research context. Fitting methods are discussed and compared, and the resultant fits to the datasets of interest are examined across the families. Generally speaking, it is found that numerical maximum likelihood estimation is both reasonably reliable and, although time-consuming, computationally robust. However, there are indications that this robustness may not apply across the whole parameter space. In addition, quick methods involving quantile-based skewness and kurtosis functionals provide good indicators of distributional suitability as well as sound starting points for numerical procedures.

Similarities and contrasts in comparisons of the resultant fits to the datasets of interest lead to investigations to directly compare the theoretical properties of the distributional families. These comparisons make use of the general theoretical framework describing skewness and kurtosis.

The results and relationships that are revealed have opened a number of pathways for further research. There is the ongoing challenge of extending coverage of different skewness and kurtosis properties that are observed in real data, and this could be investigated through the use of different bases and different transformations.

Kristen Gibbons *currently works as the Data Management and Analysis Team Leader and Statistician within the Clinical Research Support Unit, Mater Medical Research Institute in Brisbane. Her primary work involves being a collaborative researcher, consultant and educator of doctors, nurses, scientists and allied health staff in a medical research environment. Her areas of interest include educating clinicians and researchers about the appropriate use of statistics in medical research, as well as investigating statistical models for customizing birthweight.*

DOES POPULATION DENSITY AFFECT RABBIT SURVIVAL WHERE RABBIT HAEMORRHAGIC DISEASE VIRUS (RHDV) IS ENDEMIC?

Khageswor Giri[1], Kym L. Butler[1], Esther Meenken[2], Hwan-Jin Yoon[3], and Steve Mcphee[1,4]

[1]Future Farming Systems Research Division, Department of Primary Industries
600 Sneydes Road, Werribee, Victoria, 3030, Australia
kym.butler@dpi.vic.gov.au
khageswor.giri@dpi.vic.gov.au

[2]Plant & Food Research
Private Bag 4704, Lincoln, N Z
meenkene@crop.cri.nz

[3]School of Mathematical Sciences (MSI)
Australian National University
Canberra ACT 0200, Australia
Hwan-jin.yoon@anu.edu.au

[4]Agricultural Technical Services
177 Ballan Road, Werribee, Victoria 3030, Australia
steven.mcphee@bigpond.com

Rabbit haemorrhagic disease virus (RHDV) is endemic in Victoria. In epidemiological studies, it is a commonly held view that the spread of disease is density dependent. It was hypothesised that the impact of RHDV should also be density dependent. The impact of RHDV potentially depends on the number of previously infected animals surviving and/or the number of susceptible animals in the population. A five year capture-recapture study was carried out in Bacchus Marsh (Victoria) from year 1998 to year 2003. The effect of population density on rabbit survival was examined within seasons and age classes by using Cormack-Jolly-Seber models. The intricacies of selecting different categories of rabbits while examining the effect of population density on rabbit survival will be discussed. Contrary to popular belief, there was no indication of an effect of population density indicators on rabbit survival.

Khageswor Giri *joined Department of Primary Industries in 2009 immediately after completing a PhD in Statistical Science from La Trobe University. He is a member of the Statistical Society of Australia and the Golden Key International Honour Society. Khageswor enjoys his work with capture-recapture models and has also been providing statistical consulting to Fisheries Victoria.*

MINMAXENT MODELLING IN TIME SERIES

Aladdin Shamilov[1], Cigdem Giriftinoglu[2]

[1] Anadolu University, Faculty of Science, Department of Statistics 26470 Eskişehir, Turkey
asamilov@anadolu.edu.tr

[2]Anadolu University, Faculty of Science, Department of Statistics 26470 Eskişehir, Turkey
cgiriftinoglu@anadolu.edu.tr

In this study, by starting from a Maximum Entropy (MaxEnt) distribution for a stationary time series, a special class of distributions called MinMaxEnt Distributions is defined. It is proved that entropies of these distributions form a monotonic decreasing sequence bounded below by the entropy of the given distribution. In other words, as the number of autocovariances increases, the entropy of MinMaxEnt distribution goes on decreasing. This property of MinMaxEnt distributions allows us to achieve the best modelling in time series. This modelling is applicable to many statistical problems. We apply MinMaxEnt modelling to real industrial time series data using a MATLAB program.

References

Kapur, J.N. and Kesavan, H.K. (1992). Entropy Optimization Principles with Applications. USA: ACADEMIC PRESS, INC.

Wei, W.S.(2006). Time Series Analysis, Univariate and Multivariate Methods. USA:Pearson Education, Inc.

Pourahmadi M. and Soofi E., (1998). Prediction variance and information worth of observations in time series. Journal of Time Series Analysis, 21, 4, 413-434.

Cigdem Giriftinoglu *graduated from the Department of Statistics of Anadolu University in 2001. Cigdem finished her Masters Degree in Statistics in 2005. She completed her PhD in 2009 and currently works as a research assistant in Anadolu University Department of Statistics.*

DATA PREDICTION COMPETITIONS: MORE THAN JUST A BIT OF FUN

Anthony Goldbloom

Kaggle Pty Ltd
anthony.goldbloom@kaggle.com

Kaggle is a global platform for data prediction competitions allowing companies to post their problem and have it scrutinised by the world's statisticians and computer scientists. By exposing a problem to a wide range of analysts and techniques, data prediction competitions turn out to be a great way to get the most out of a dataset, given its inherent noise and richness. For example, Kaggle has been running a bioinformatics competition requiring participants to pick markers in HIV's genetic sequence that predict a change in viral load (a measure of the severity of infection). Within a week and a half, the best submission had already outdone the best methods in the scientific literature.

Anthony Goldbloom *is the Founder and CEO of Kaggle Pty Ltd, a global platform for data prediction competitions. In addition to founding and building Kaggle, Anthony continues to consult to hosts of Kaggle competitions to help them frame modelling tasks, to get the best out of the new platform. Before Kaggle, Anthony was a macroeconomic modeller for the Reserve Bank of Australia and before that the Australian Treasury. In these roles, Anthony built and maintained macroeconomic models of Australia's economy, helping to improve forecasting and to model the economic effect of changes in policy parameters. Anthony graduated with first class honours in econometrics at the University of Melbourne. He has published in The Economist magazine and the Australian Economic Review.*

MERGING SURVEILLANCE DATA- AN EXPERIENCE REPORT

[1]Norm Good, [2]John O'Dwyer, [3]Christine O'Keefe

[1]CSIRO Mathematics, Informatics and Statistics/ Australian e-Health Research Centre
Level 5, UQ Health Sciences Building 901/16Royal Brisbane and Women's Hospital
Herston, QLD, 4029
Norm.Good@csiro.au

[2]Australian e-Health Research Centre, CSIRO
Level 5, UQ Health Sciences Building 901/16Royal Brisbane and Women's Hospital
Herston, QLD, 4029
John.O'dwyer@csiro.au

[3]CSIRO Mathematics, Informatics and Statistics
Building 108, Australian National University, Acton ACT 2601
Christine.O'keefe@csiro.au

The MSD project brings together two colorectal cancer (CRC) surveillance datasets. One from a hospital in Melbourne, and one from two hospitals in Adelaide. The main objectives of the project are to verify the use of Faecal Occult Blood Tests as a predictor of CRC; determining the optimal interval between colonoscopies within a surveillance framework; and determining the impact of family history as a predictor for CRC.

Each hospital collects the same information on each patient, however there are significant differences in philosophy and semantics. This mainly reflects the particular research interests of the lead clinicians. Two main challenges became apparent: Firstly, to create an integrated dataset whereby values of the individual variables are consistent; Secondly, understanding the processes that created the data.

In this talk we will highlight the approach that we took to identify and solve these issues to create and analyse a consistent and coherent dataset.

Norm Good *currently works as a Statistician with the Australian E-Health Research Centre, Herston, in Queensland. His current area of research is longitudinal data analysis and the analysis of linked health datasets. Norm is also a statistical consultant with the group, advising on variable selection techniques, study design, survival analysis and statistical computing. His previous position was as a Stock Assessment Modeller with the QLD Dept, Primary Industries and Fisheries.*

A SUITE OF VARIABLE SELECTION METHODS TO IDENTIFY A SHORTLIST OF PROTEIN BIOMARKERS

Doecke J.[1], Buckley M., Dunne R., Good, N. [2], Wilson, W.

The Australian E-Health Research Centre, CSIRO Preventative Health Flagship,
CSIRO Mathematics, Informatics and Statistics
Level 5 UQ Health Sciences Building (901/16), Royal Brisbane & Women's Hospital
Herston, Queensland 4029, Australia
[1]James.doecke@csiro.au
[2]Norm.Good@csiro.au

We developed an analysis pipeline from publicly available statistical methods to take raw protein array data, and select a panel of markers for class prediction. We composed an iterative loop that randomly selected 70% of the complete data set (training data), computed multiple imputations (Multivariate Normal Imputation (MVNI) and Multiple Imputation by Chained Equations (MICE)), and ran statistical models (Random Forest (RF), Generalized Boosted Models (GBM), Classification Trees (CT) and Linear Models for MicroArray (LIMMA)) 100 times to select a shortlist of biomarkers. The top 20% of biomarkers were chosen from each iteration of the loop, and the average top 20% chosen from each method. Biomarkers that were chosen in all four statistical methods were selected for prediction using General Linear Model (GLM) and Receiver Operating Characteristic (ROC) analyses. All statistical analyses were performed in R. After using two separate multiple imputation methodologies, we found no statistical evidence that one performed better than the other. We show biomarker agreement across four separate statistical methods to separate the two classes. We find that while Random Forest (RF) is very robust in producing repeatable results over multiple conditions, Generalized Boosted Models (GBM) appears highly sensitive to changes in input parameters. Linear Discriminant Analyses results are highly reproducible over a variety of parameters, however provides lower accuracy estimates than both RF and GBM. As a prediction tool, we found the use of the ROC method useful for the calculation of accuracy statistics. These analyses were performed on data that was able to be transformed to fit a normal distribution, and as such the multiple imputation methods MVNI and MICE were easily interchangeable. Other imputation methods should be considered where the data does not conform to Gaussian standards.

References

R: A Language and Environment for Statistical Computing, R Development Core Team, R Foundation for Statistical Computing, Vienna, Austria.

Norm Good *currently works as a Statistician with the Australian E-Health Research Centre, Herston, in Queensland. His current area of research is longitudinal data analysis and the analysis of linked health datasets. Norm is also a statistical consultant with the group, advising on variable selection techniques, study design, survival analysis and statistical computing. His previous position was as a Stock Assessment Modeller with the QLD Dept, Primary Industries and Fisheries.*

THE CURIOUS TAIL OF *PREP*

Ian Gordon[1], *Ray Watson*[2]

[1] *Statistical Consulting Centre, The University of Melbourne Victoria, Australia 3010*
irg@unimelb.edu.au

[2] *Department of Mathematics and Statistics, The University of Melbourne, Victoria, Australia 3010*
ray.watson@unimelb.edu.au

Motivated by dissatisfaction with null hypothesis testing and abuse of the *P*-value in particular, Killeen (2005) proposed a new way to represent an inference about a single parameter, called the "probability of replication", or p_{rep}. In this definition, replication was defined to be an effect of the same sign as that found in the original experiment. The introduction of this measure was enthusiastically welcomed as a revolution: this "may change how all pyschologists report their statistics" (Cutting, 2005), and it has indeed been used in the simple representation of inferences in published papers, as an alternative to the *P*-value or a confidence interval.

We show that the derivation originally offered is wrong, and demonstrate that a correct implementation of the basic idea leads to a measure which is simply 1 minus the *P*-value; hence the "curious tail" of the title.

The story of the introduction and impact of p_{rep} is an interesting study in itself; we provide an account of this and comment on the implications for publication processes.

References
Cutting, J.E. (2005). Acknowledgement. Psychological Science, 16, 1013.
Killeen, P.R. (2005). An alternative to null-hypothesis significance tests. Psychological Science, 16, 354-353.

Ian Gordon *is the Director of the Statistical Consulting Centre at the University of Melbourne. He has worked as a statistical consultant on projects from across the whole range of statistical applications over the last 25 years, and also has a strong involvement and interest in statistical education. He is currently the President of the Victorian Branch of the Statistical Society of Australia.*

OPEN DISCUSSION FORUM FOR SSAI BIOSTATISTICS SECTION ACTIVITIES

Mark Griffin

Co-Chair of the SSAI Biostatistics Section

The last two years have seen a lot of activity within the SSAI Biostatistics Section. Major events include the establishment of a six-monthly visiting workshop presenter program. This program has seen a workshop on Causal Inference by Liliana Orellana on the 4th and 5th of September, and a workshop on Hot Topics in Clinical Trials by Scott Evans and Rui Wang immediately preceding the ASC2010 conference. The Section also organized and presented a two-hour seminar session at the Joint Statistical Meeting in Vancouver in August, and warmly thank Ian Marschner, Niels Becker, and Val Gebski as our presenters. We are delighted to see a Biostatistics session in every time-slot of the ASC2010 program.

In this discussion forum the co-chairs of the SSAI Biostatistics Section will present a summary of the Section's activities over the last two years and our plans between now and ASC2012. We warmly welcome discussion on additional activities that the Section could be pursuing, and how the Section could better be serving our members.

Mark Griffin *is the Co-Chair of the SSAI Biostatistics Section (along with Ian Marschner). Mark currently works for the School of Population Health at the University of Queensland. He has a keen interest in developing statistical capacity within developing nations (particularly in the Pacific Islands), and is the Chair of the New Projects Committee for Statistics Without Borders. SWB is a special-interest group of the American Statistical Association dedicated to providing pro bono statistical consultancy to developing nations.*

FLAVOURS OF STATISTICS: THE DOING OF LAMIA GURDLENECK

David Griffiths

School of Mathematics and Applied Statistics
University of Wollongong NSW2522
griffd@uow.edu.au

More so than most other sciences, the development of the discipline of Statistics since 1900 has been largely driven by people from England and its former colonies, most notably the USA. But other colonies which were more recently released from the motherland's apron strings, especially Australia, Canada and New Zealand, have all punched above their weight. See, for example Peter Sprent's forthcoming article on the adolescent phase in the development of the discipline in Australia.

The contributions from the different countries reflect national culture as well as individual idiosyncrasy. This talk will focus on the contribution of one Englishman, Maurice George Kendall. Although he made numerous significant contributions to Statistics and other mathematical sciences, Kendall is best known for his mammoth (1, 2, 3 or 4 - or more - volume, depending on when you count) work, The Advanced Theory of Statistics. This talk will trace the development of one important part of Volume 2, and its relation to the man, his life and his role in the discipline.

The talk will also contrast the man and his work with that of some of his disciplinary cousins from across the Atlantic, in particular Jack Kiefer, who were less than impressed by the integrity of the Advanced Theory.

This research was facilitated by the Master and Fellows of St John's College Cambridge who gave me access to the Kendall papers (recently donated by MGK's son Peter to the college library).

David Griffiths *has been Foundation Professor of Statistics at the University of Wollongong since 1987. There was one statistician on the staff at the then Department of Mathematics when he took up the position. There are now four Statistics Professors in an active Statistical group within the School of Mathematics and Applied Statistics. After spending seven years focusing on University governance while chair of the academic Senate, David is enjoying the return to a normal academic role.*

ESTIMATION OF BETWEEN- AND WITHIN-PAIR REGRESSION EFFECTS IN LOGISTIC REGRESSION WITH SHARED MEASUREMENT ERROR

Lyle C. Gurrin[1], Elizabeth J. Williamson[2], Martin L. Hazelton[3]

[1]Centre for Molecular, Environmental, Genetic & Analytic Epidemiology
Melbourne School of Population Health
Parkville, Melbourne, Australia
lgurrin@unimelb.edu.au

[2]Department of Epidemiology and Preventive Medicine
The Alfred Centre, Monash University
99 Commercial Road, Melbourne, Victoria
ewi@unimelb.edu.au

[3]Institute of Fundamental Sciences - Statistics
Massey University, Private Bag 11222
Palmerston North, New Zealand
M.Hazelton@massey.ac.nz

Twin studies provide naturally matched pairs that can exploit within-pair comparisons of data to avoid confounding exposure-outcome associations by shared factors. For binary outcomes, paired data can be analysed using the logistic regression model of Neuhaus & Kalbfleisch (1998) with linear predictor

$$\beta_0 + \beta_w \frac{1}{2}(x_{ij} - x_{ik}) + \beta_b \bar{x}_i \ ,$$

where *i* indexes pair and $j \neq k$ index within pair. When estimates of the between- and within-pair effects β_b and β_w differ it raises questions regarding the interpretation of the estimate of the former, and whether it provides useful information about the latter.

If $\beta_b = \beta_w = \beta$ then one scenario where the issue of differing estimates of β_b and β_w has a straightforward resolution is when the pair covariate mean $\bar{x}_i$ is measured with error, but that the within-twin difference is subject to negligible error. For instance, siblings reporting nutritional intake may be accurate in comparison to each other, but less so on an absolute scale. Failure to account for the measurement errors leads to attenuation in the estimates of β_b, generating an apparent discrepancy with β_w. By using the SIMEX method (Cook & Stefanski 1994) with shared within-pair measurement error, it is possible to adjust for this and generate an estimate of β that is considerably more efficient than estimating β_w alone or using conditional logistic regression.

We examine the efficacy of this approach through a simulation study, and illustrate its use through an example exploring the association between low birth weight and cord blood erythropoietin (EPO) as a marker of hypoxic stress *in utero* and possible growth restriction (Carlin 2005). In this example $\widehat{\beta}_b$ = 0.34 (s.e.= 0.13) and $\widehat{\beta}_w$ = 0.62 (s.e.= 0.23). The SIMEX-adjusted estimate of β in a single-parameter model is 0.49 (s.e.=0.13), closer to the estimate of β_w than the unadjusted estimate $\widehat{\beta}$ = 0.40 (s.e.=0.12) but with similar precision.

References
Neuhaus JM & Kalbfleisch JD (1998). Between- and within-cluster covariate effect in the analysis of clustered data. Biometrics, 54, 638 - 645.
Cook JR & Stefanski LA (1994). Simulation-extrapolation estimation in parametric measurement error models. Journal of the American Statistical Association, 89, 1314 –- 1328.
Carlin JB, Gurrin LC, Sterne JAC, Morley R, Dwyer T (2005). Regression models for twin studies: a critical review. International Journal of Epidemiology, 34, 1089 –- 1099.

Lyle Gurrin *is Associate Professor of biostatistics in the Centre for MEGA Epidemiology at the Melbourne School of Population Health (MSPH). Until 2001 he worked in medical research in Western Australia, during which he completed a PhD in biostatistics at the Institute for Child Health Research, worked as the sole consulting biostatistician for Princess Margaret and King Edward Memorial Hospitals and collaborated on both basic science and epidemiological programs studying the developmental origins of health and disease. Dr Gurrin is now the co-principal investigator of the NHMRC-funded HealthIron study of genetic and environmental modifiers of hereditary haemochromatosis, and is a co-investigator of several large NHMRC-funded cohort studies of allergic diseases based in Melbourne. He teaches postgraduate coursework students in public health and epidemiology through the MPH and MEpi programs at the MSPH, and in biostatistics through his involvement with the Biostatistics Collaboration of Australia (BCA).*

HYPEREXTENDED AGE PERIOD COHORT ANALYSIS OF (AGE, PERIOD)-TABULATED DATA ON BREAST CANCER DEATHS FOR FEMALES IN THE US AND JAPAN

Nobutane Hanayama

Shobi University
655 Shimomatsubara, Kawagoe 3501153, Japan
nob-hanayama@jcom.home.ne.jp

In the field of breast cancer epidemiology, in addition to the age effect, there has been an interest in the environmental effect on the disease associated with behavior, diet, and physical and chemical exposures. Meanwhile, the period effect, such as one associated with advancing medical technologies, is indisputable. So, in this study, for projecting such age, period and environmental risks on breast cancer, a hyperextended age period cohort (HAPC) model is applied to the analysis of (age, period)-tabulated data on breast cancer deaths for females in the US and Japan. It is shown that the HAPC model is free from the non-identifiability problem from which the original age, period cohort (APC) model suffers. Further it is seen that the HAPC model provides a better fit than the APC model on data for females in the US and for females in Japan in terms of Akaike's Information Criteria. In addition most likelihood estimates of the age, period and environmental effects on the disease suggest that they have roughly common trends for both countries even though the ethnic compositions, the life environments and the government policies for public health are different between the countries. Finally, it is concluded that the environmental risk for the disease has been rapidly decreasing since the 1980s in both countries.

References

Hanayama, N. (2001). A simple two-stage model for cancer risk in the environment. Environmetrics 12, 757-773.
Hanayama, N. (2004). Age-environment model for breast cancer. Environmetrics 15, 219-232.
Hanayama, N. (2007). An extended age period cohort model for analysing (age, period)-tabulated data. Statistics in Medicine 26, 3459-3475.

Nobutane Hanayama *currently works as a associate professor of Shobi University in Kawagoe, Japan. His area of interest is the biostatistics or epidemiology. He is also involved with several projects concerning regional development as a questionnaire analyst.*

A FAMILY OF ESTIMATORS FOR SINGLE AND TWO-PHASE SAMPLING USING TWO AUXILIARY ATTRIBUTES

Muhammad Hanif[1], *Inam ul Haq*[2]

[1]*Lahore University of Management Sciences, Lahore Pakistan*
hanif@lums.edu.pk

[2]*National College of Business Administration &Economics*
Lahore, Pakistan
inam-ul-haq786@hotmail.com

Jhajj et al. (2006) proposed a general family of estimators and derived a general expression for mean square error of these estimators by using a single auxiliary attribute. Hanif et al. (2009) proposed a more general form of the Jhajj et al. (2006) family of estimators and derived an expression for mean square error of these particular estimators in the proposed family by using a "k" auxiliary attribute. In this paper we have suggested some new estimators. We have also derived the mean square error expression of Jhajj et al. (2006) in the case of partial information, using two auxiliary attributes. Mathematical comparisons of these estimators have been made. Empirical study has also been conducted to show that the new estimators are more efficient. Moreover full information cases are more efficient than partial and no information cases.

AGRI-ENVIRONMENTAL INFORMATICS: THE CONTRIBUTION OF STATISTICAL SCIENCE AND STATISTICIANS

Bronwyn D. Harch

CSIRO Mathematics, Informatics & Statistics & Sustainable Agriculture National Research Flagship
Level 3, 306 Carmody Road, St Lucia, QLD 4067, AUSTRALIA
Bronwyn.Harch@csiro.au

Agri-Environmental Informatics research is focused on a fresh look at agricultural productivity and agricultural futures. Research investment in Agri-Environmental Informatics is to ensure innovations in "smarter information use" contributes to reducing the carbon footprint of Australia's land use, whilst also achieving the productivity gains needed for prosperous agricultural and forest industries and global food security.

Why is this important? Agri-Environmental Informatics is a critical underpinning set of information and methodologies for Australia's future management of carbon, agri-ecosystem health, agricultural productivity and the nation's wellbeing and prosperity. Government and industry are seeking more knowledge, reliable data and expertise to help them understand the competing and conflicting pressures on our agri-ecosystem balance. This research focus will provide the underpinning information systems to aid evidence based decision making. Investment is aimed at ensuring Australia maintains a leadership role in the contribution of agri-environmental systems, analytical methodologies and spatially enabled sensing technologies that will facilitate globally consistent information assets, which are also locally relevant to Australian agriculture.

This paper will outline the contribution statistical science can make to this emerging research area and importantly the role of the statistician as an integrator in multidisciplinary research teams.

Bronwyn Harch *currently works as Deputy Director of CSIRO Sustainable Agriculture National Research Flagship, in Brisbane. Her area of research interest has been on the statistical design of landscape scale sampling protocols and monitoring programs, as well as the statistical modelling of complex landscape systems. She has led a number of major statistical projects within large multidisciplinary landscape-based studies. For Bronwyn - "informatics" is the science of turning data into insight and action.*

IMPROVING THE EFFICIENCY OF PIXEL-BASED REGRESSION FOR SPATIAL POISSON POINT PROCESSES

Adrian Baddeley[1], Andrew Hardegen[2], Robin K. Milne[3]

[1] CSIRO Mathematics, Informatics and Statistics
Private Bag 5, Wembley WA 6913
Adrian.Baddeley@csiro.au

[2] School of Mathematics and Statistics, M019, University of Western Australia
35 Stirling Highway, Crawley WA 6009
hardegen@maths.uwa.edu.au

[3] School of Mathematics and Statistics, M019 University of Western Australia
35 Stirling Highway, Crawley WA 6009
milne@maths.uwa.edu.au

Spatial point pattern data are usually analysed by point process methods not requiring any discretisation of the data. However, in some other fields, investigators analyse such data by partitioning the observation window into relatively small pixels. Representative or 'surrogate' covariate values are then assigned to each pixel and a generalised linear model (GLM) is fitted to the discretised observations, i.e. pixel counts or presence/absence indicators.

For increasingly fine pixel grids, the method which is often termed spatial logistic regression is asymptotically equivalent to a Poisson point process with loglinear intensity. Despite this, spatial logistic regression does not correspond to any point process model in continuous space, and the regression parameters carry physical meaning only for very fine pixel grids.

We assume that the observed point pattern is a realisation of a Poisson point process with loglinear intensity. If the covariate function is constant within each pixel, then the discretised observations conform exactly to a GLM: loglinear Poisson regression for counts, or complementary log-log regression for indicators (rather than logistic regression!). In general, however, the observations do not conform to any GLM. Nevertheless, a GLM may be fitted by choosing surrogate covariate values for each pixel; both logistic regression and complementary log-log regression then yield practical approximations. Examples suggest, however, that the biases caused by surrogate approximation are often much more severe than those caused by incorrect choice of link function.

The `pragmatic' solution has been to employ extremely fine pixel grids, which is computationally expensive. Baddeley et. al. (2010) outline two possible strategies for improving the efficiency of pixel-based regression while simultaneously containing computational load. One of these, which we call `pixel splitting', is new. We describe this strategy, and point towards a general theory of performance.

References

Baddeley, A., Berman, M., Fisher, N.I., Hardegen, A., Milne, R.K., Schuhmacher, D. and Shah, R. (2010). Spatial logistic regression and change-of-support in Poisson point processes. Electronic J. Statistics, under revision.

Andrew Hardegen *is completing a PhD at the University of Western Australia about efficient parameter estimation for spatial point process models. He has wide interests in statistical theory and point processes.*

MODELLING NON-STATIONARY SPATIAL COVARIANCE BY SPECTRAL TEMPERING, WITH APPLICATION TO NITROUS OXIDE EMISSIONS FROM SOIL

Kathy Haskard[1], Murray Lark[2], Sue Welham[2]

[1] Australian Mathematical Sciences Institute (AMSI)
111 Barry St, c/o University of Melbourne, Vic 3010, Australia
kathy@amsi.org.au

[2] Rothamsted Research
Harpenden, Hertfordshire, AL5 2JQ, UK
murray.lark@bbsrc.ac.uk

Spectral tempering for modelling non-stationary spatial covariance is outlined. It can be applied to single-realisation spatial data in one or more spatial dimensions, does not require regularly-spaced observations, and guarantees positive definite covariance. Autocorrelation and variance of the spatially-correlated random effects, and the nugget variance, are independently modelled as functions of continuous or discrete auxiliary variables, including location, pH, and soil classes. The method is set within a linear mixed models framework, with parameters estimated by REML. In application to a nitrous oxide soil emissions data set, non-stationary models fit the data better than do the commonly-assumed stationary models, and validation shows that they yield more realistic estimates of prediction error variance (kriging variance).

References

Haskard, K.A., Welham, S.J., and Lark, R.M. (2010). Spectral tempering to model non-stationary covariance of nitrous oxide emissions from soil using continuous or categorical explanatory variables at a landscape scale. Geoderma 159: 358-370.

Kathy Haskard *has been a statistical consultant in several organisations in the Australian public sector for over 25 years, mainly working in agricultural, biological, environmental and fisheries areas. She completed a PhD in 2007, using linear mixed models for model-based geostatistics, with research into development and estimation of the Matérn spatial correlation function extended to allow anisotropy. She then completed a three-year post-doctoral appointment at Rothamsted Research in the UK, working with Murray Lark on modelling non-stationarity in spatial covariance functions. Recently she was appointed as the inaugural statistician in the Australian Mathematical Sciences Institute (AMSI), aiming to promote the use and understanding of the mathematical sciences across Australia, via programs in education, science and industry, in conjunction with half her time spent with Parks Victoria.*

ON ESTIMATION OF LOGARITHMIC AUTOREGRESSIVE CONDITIONAL DURATION (LOG-ACD) MODELS WITH GENERALIZED GAMMA ERRORS USING ESTIMATING FUNCTIONS: A SIMULATION STUDY

Ng Kok Haur[1], Shelton Peiris[2], David Allen[3]

[1]Institute of Mathematical Sciences, Faculty of Science, University of Malaya
50603 Lembah Pantai, Kuala Lumpur, Malaysia
kokhaur@um.edu.my

[2]School of Mathematics and Statistics F07, University of Sydney, NSW 2006, Australia
shelton@maths.usyd.edu.au

[3]School of Accounting, Finance and Economics, Edith Cowan University, WA
d.allen@ecu.edu.au

It is known that the Generalized Gamma (GG) distribution is more flexible than many other distributions in financial applications. This paper considers the estimation of Log-ACD models with errors from a GG using the theory of estimating functions (EF). Using a large scale simulation study based on GG Log-ACD models, it is shown that the EF approach is easy to apply and provides comparable estimates with the maximum likelihood method.

References

Bauwens,L., Giot,P., The Logarithmic ACD Model: An Application to the Bid-ask Quote Process of Three NYSE Stocks. Annales D'Economie et de Statistique. 60(2000) 117-145.

Godambe, V. P., The Foundations of Finite Sample Estimation in Stochastic Processes. Biometrika. 72:2 (1985) 419-428.

Pathmanathan,D., K.H.Ng., Peiris, M.S., On Estimation of Autoregressive Conditional Duration (ACD) Models based on Different Error Distributions, Sri Lankan Journal of Applied Statistics. 10 (2009) 251-269.

Ng Kok Haur *obtained his Ph.D. from the University of Malaya in 2006. Currently, he is a senior lecturer at the University of Malaya, Kuala Lumpur, Malaysia. His research interests are regression analysis and financial time series.*

SURVIVAL STATUS OF OESOPHAGEAL CANCER PATIENTS IN ASSAM, INDIA

Jiten Hazarika[1], Manash Pratim Barman[2], Rup Kumar Phukan[3]

[1]Department of Statistics, Dibrugarh University, Assam
Reader, Department of Statistics, Dibrugarh University, Dibrugarh – 786004, Assam, India
Jitenhazarika@rediffmail.com

[2]Department of Statistics, Dibrugarh University, Assam
Lecturer, Department of Statistics, Dibrugarh University, Dibrugarh – 786004, Assam, India

[3]Regional Medical Research Center (ICMR), Dibrugarh
Scientist-C, Regional Medical Research Centre (ICMR) Dibrugarh, Assam, India

Background: The incidence and mortality of oesophageal cancer patients are high in Assam, India. No Biostatistical study had been conducted up till now to assess the survival pattern of oesopahgeal cancer patients in the state. This research work was undertaken to study the survival pattern of the oesophageal cancer patients with respect to some factors relevant to this region.

Methods: An analytic study was conducted using an historical cohort and information from the medical charts of patients with oesophageal cancer in Assam Medical College Hospital (AMCH), Assam. Out of 233 patients diagnoised between 1st January 2004 and 31st December 2005 in AMCH, 178 patients were followed for whom addresses were available. The data were analysed using the Kaplan Meier product limit estimator, the log-rank test and the Cox regression model.

Results: The average survival time was estimated to be 10.33 months (95% C.I. 7.86 to 12.88). The survival of patients belonging to the lower socio-economic group was worst with average survival time of 5.53 months (95% C.I. 4.373 – 6.694) and adjusted hazard ratio of 2.79 (95% C.I. 1.30-5.98) in comparison to a higher socio-economic group. About 13% of the patients remained untreated. Their average survival time was only 3.33 months (95% C.I. 1.82 – 4.846), significantly less than the treated patients. The adjusted hazard ratio of the untreated patients was 4.54 (95% C.I. 2.25 – 9.16). Patients treated in Assam experience a 2-fold risk of dying in comparison to the patients treated in metropolitan India.

Conclusion: The survival of patients with oesophagus cancer was low in Assam. Socio-economic status, treatment and stage of cancer were important factors influencing the survival of patients with oesophageal cancer. The performance of health services with regard to cancer care in Assam seemed to be poor in comparison with other states of India.

Jiten Hazarika *currently works as a Reader in the Department of Statistics, Dibrugarh University. His area of interest is Biostatistics and Econometrics. Hazarika has been involved in teaching and research in the fields of Biostatics, Epidemiology and Econometrics. He has guided four Ph.d scholars. He has published 25 research papers in different national and international journals.*

INFERENCE FOR DYNAMIC TRAFFIC MODELS

Martin L. Hazelton[1], *Katharina Parry*[2]

[1] *Massey University, Private Bag 11-222, Palmerston North, New Zealand*
m.hazelton@massey.ac.nz

[2] *Massey University, Private Bag 11-222, Palmerston North, New Zealand*
k.parry@massey.ac.nz

There is currently burgeoning interest in the transportation literature in the development of models to describe the day-to-day evolution of traffic flows over a network. We examine the problem of estimating the parameters of such models based on daily traffic counts on a subset of network links. A critical difficulty is that such link count data generally supply only indirect information about many of the parameters of interest, because the linear system that relates link flows to latent path flows is underdetermined. In principle we can apply MCMC algorithms that sample from the set of route flows consistent with the observed link flows, but in practice enumeration of this set will typically be computationally infeasible. By using a simple Markov model for traveller behaviour (conditional on link counts) we show that it is nevertheless possible to construct a suitable proposal distribution for MCMC sampling when the underlying network has certain properties.

Martin Hazelton *holds the Chair of Statistics at Massey University in Palmerston North, New Zealand. Prior to that appointment he spent nine years in the School of Mathematics and Statistics at the University of Western Australia. His current research interests include modelling and inference for transport networks, and smoothing methods (particularly with applications in medicine and epidemiology).*

DESIGNING EFFICIENT BUT UNBALANCED CHOICE EXPERIMENTS

John Henstridge[1], Donna Hill[2]

[1] Data Analysis Australia
97 Broadway, Nedlands, 6009, Western Australia
john@daa.com.au

[2] Data Analysis Australia
97 Broadway, Nedlands, 6009, Western Australia
donna@daa.com.au

Choice experiments have the aim of quantifying preferences for various options, usually by estimating a monetary value that consumers might attach to each option. They are particularly attractive in transport studies where the estimates might then provide direct guidance on setting fares and determining likely patronage, hence providing information critical to cost-benefit and feasibility studies. However choice experiments often lead to large designs with many choice sets to achieve balance. While this might be possible with an online survey where questionnaires can be generated as required, it is difficult to implement in other circumstances. In addition, many balanced designs can lead to comparisons being presented that are either not meaningful or not challenging.

In designing a study for a proposed new rail link, the implementation constraints were extreme – several interview modes were used, cost constraints limited the number of distinct choice sets and one option – no new link – was so dominant in peoples minds that it had to be included at most times. This made it impossible to have an ideal design. A search procedure was therefore used to generate an almost optimal design within these constraints. The search used both a random approach to prevent being locked into local maxima as well a local improvement. The result was a design that was successfully implemented and provided valid estimates of customers' values on the various options associated with the rail proposal.

John Henstridge *is the founder and Managing Director of Data Analysis Australia, a consulting company in statistics and mathematics. John's work was originally in time series and statistical computing but as a consultant he has worked in many other fields as well.*

ROBUST METHODS IN BIOSTATISTICS

Stephane Heritier

The George Institute, the University of Sydney
PO Box M201, Missenden Road, NSW 2050
sheritier@george.org.au

Robust statistics is an extension of classical statistics that specifically takes into account that the underlying models used by analysts are only approximate. The basic philosophy of robust statistics is to produce statistical procedures that are stable with respect to small changes in the data or model departures. Such methods are now readily available for most models commonly used in biostatistics including models for repeated measures and survival data. As an introduction to this topic, the talk will focus on robust estimation and inference for generalized linear models and/or longitudinal data analysis via generalized estimating equations (GEE). Applications to real data will be given to illustrate the interest and feasibility of such analyses.

References

S. Heritier, E. Cantoni, S. Copt and M.P. Victoria-Feser (2009) Robust Methods in Biostatistics, Wiley: Chichester, UK.

Stephane Heritier *is Associate Professor at the University of Sydney. He works as Head of Statistical Research at the George Institute for Global Health. He has been involved in designing and analysing many trials and epidemiological studies, teaching biostatistics (Master of Biostatistics, Biostatistics Collaboration of Australia) and consulting activities. His current research interests include adaptive designs, robust methods in biostatistics, cluster-randomised trials, multistate models and applications of saddlepoint techniques to medical data.*

INCORPORATING RISK IN STRATEGIC DECISION MAKING PROCESSES

Donna Hill[1], Dr John Henstridge[2], Prudence Thompson[3]

[1]Data Analysis Australia
97 Broadway Nedlands WA 6009
donna@daa.com.au

[2]Data Analysis Australia
97 Broadway Nedlands WA 6009
john@daa.com.au

[3]Data Analysis Australia
97 Broadway Nedlands WA 6009
prudence@daa.com.au

While statisticians realise the importance of taking into account the uncertainty of predictions when presenting model results, it can be difficult to convince non-statisticians of this. However, occasionally Data Analysis Australia is approached by clients who understand that the uncertainty of model predictions are a measure of risk that they need to incorporate into their strategic decision making processes. For a national agricultural investment company, statistical models were developed to forecast likely distributions of crop values, based on distributions of historical yield and prices. The key focus of the models was obtaining a good understanding of the variance of the distributions. For an energy provider, we reviewed the modelling approach they had employed to obtain realistic predictions of the benefits of a set of improvement strategies and revealed that there was some risk of overestimating the benefits. For a third client, we implemented a simulation approach to help them understand the risk involved in using a single year's data to estimate annual peaks.

Donna Hill *is a Senior Consultant Statistician at Data Analysis Australia, Perth. She has managed or worked on over 200 client projects in her eleven years at Data Analysis Australia. Donna has specific expertise in the areas of statistical modelling, surveys, road safety, transport and mining.*

DEVELOPING SMALL AREA ESTIMATES FOR A STATE HEALTH SURVEY

Diane Hindmarsh[1], David Steel[2], Ray Chambers[3], Margo Barr[4]

[1]Centre for Statistical and Survey Methodology
University of Wollongong NSW 2522
dmh972@uow.edu.au

[2]Centre for Statistical and Survey Methodology
University of Wollongong NSW 2522
dsteel@uow.edu.au

[3]Centre for Statistical and Survey Methodology
University of Wollongong NSW 2522
ray@uow.edu.au

[4]Centre for Epidemiology and Research
NSW Department of Health, Locked Bag 961, North Sydney NSW 2059
mbarr@doh.health.nsw.gov.au

Over the past 25 years a strong theoretical framework has been developing around small area estimation methods, using both unit-level and area-level models. For the estimates to be used in a practical setting there needs to be clear understanding of the proposed use of the estimates and rigorous assessment of the reliability and robustness of the small area estimates calculated from the survey source. End users of small area estimates may not necessarily understand the theoretical background of the methods, but they need sufficient background information to make informed choices about the applicability of the estimates.

NSW Health has collected data on health risk factors and health status through a continuous population health survey since 2002. The survey was designed to provide estimates for the state and the 8 health areas, with a sample size of approximately 1000 observations in any year from each of the health areas. However there is an increasing desire for estimates at lower levels of geography, in particular at the local government area level where there are just over 150 in NSW, ranging in population from over 250,000 to less than 5,000. This paper will showcase the results of a number of SAE methods as applied to the NSW population health survey data, together with the methods being used to test the reliability and robustness of the estimation methods. We will also focus on some of the issues faced when applying SAE methods to an ongoing survey.

Diane Hindmarsh *is currently undertaking postgraduate studies in small area estimation methods at the Centre of Survey Methodology at the University of Wollongong.. Prior to this Diane was employed on the Biostatistical Officers Training Program run by the NSW Department of Health.*

SPATIAL DIFFERENCES IN LYMPH NODES OF BREAST CANCER AND HEALTHY PATIENTS

Susan Holmes[1], Nelson Ray[2]

[1]Statistics Department, Sequoia Hall, Stanford
susan@stat.stanford.edu

[2]Statistics Department, Sequoia Hall, Stanford
ncray@stanford.edu

Through new statistical image segmentation software we have developed (GemIdent), we are able to identify, localize and classify Tcells, Dendritic cells, cancer cells and B cells in images of lymph nodes of cancer patients.

Spatial data analyses using Ripley's K function and Moran's I statistic have enabled us to detect changes in distributions of Tcells and Bcells in the sentinel lymph nodes and link this through survival analyses to patient prognosis.

We show how methods developed to detect clustering in population studies can be effectively used for cell populations.

References

Holmes, S.P., Kapelner, A. & Lee, P.P. (2009). An interactive JAVA statistical image segmentation system: GemIdent. Journal of Statistical Software, 30(10).
Kohrt, H.E., Nouri, N., Nowels, K., Johnson, D., Holmes, S. and Lee, P. (2005). Profile of Immune Cells in Axillary Lymph Nodes Predicts Disease-Free Survival in Breast Cancer. PLoS Medicine 2(9).
Setiadi AF, Ray, NC, Kohrt HE, Kapelner A, Carcamo-Cavazos V, Levic EB, Yadegarynia S, van der Loos CM, Schwartz EJ, Holmes S, Lee PP. (2010) Quantitative, architectural analysis of immune cell subsets in tumor-draining lymph nodes from breast cancer patients and healthy lymph nodes. PLoS One. 2010; 5 (8).

Susan Holmes *is Professor of Statistics at Stanford University, California, USA. Susan has been involved in studying the interactions between the immune system and cancer for 9 years. She teaches courses on Data Mining, Multivariate Statistics, the Bootstrap and Nonparametric statistics.*

THE STATISTICS OF EVIDENCE: APPLYING A RISK-ETHICS FRAMEWORK TO POLICY SETTING

Stephen Horn

Dept of Families, Housing, Community Services and Indigenous Affairs
Tuggeranong Office Park, Athllon Drive, Greenway ACT
stephen.horn@fahcsia.gov.au

The discipline of statistics is applied routinely to situations of imperfect prior knowledge – whether of that arising from complexity in causes (analysis) or in aggregative or projective mechanisms (inference). Policy is developed typically by employing partial knowledge within known constraints to well defined ends. There is much value placed on basing policy on evidence - that is on statistical analysis of objectively collected facts, as distinct from theory, ideology or prejudice.

But the field of study and policy position adopted (whether on a basis of objective knowledge and scientific inference – roughly what is conveyed by the expression evidence in this context - or not) are not independent. Policy is about transforming something inchoate, such as the general health of the population, into a shaped framework for governmental action – program expenditure, regulations on providers, allocation of capital funds. The state of health is a consequence of policy as well as the ground from which policy is developed. In this a policy maker is both an actor and an interpreter (to critical outsiders) of the field, so a transparent approach to turning evidence into policy positions and debate is called for. This calls for engagement on behalf of the discipline and profession of statistics in what is being committed in its name.

A risk-ethics framework is introduced to guide the treatment of evidence in a policy setting. It is illustrated by simple examples.

References

Asher, J., Banks, D., Scheuren, F. J. (eds) (2008) Statistical Methods for Human Rights. New York: Springer

Glazer, W. (ed) (2002) Rich and Poor: Disparities, Perceptions, Concomitants. Dordrecht: Kluwer Academic Publishers

Stephen Horn *works for the Australian Government as a statistician. He is interested in the application of statistics to problems of policy. As a survey methodologist and as a student of public policy he has contributed to poverty measurement, and to standards and methods for household income and expenditure collection.*

A STATISTICAL METHODOLOGY FOR ORDINAL DATA IN META-ANALYSIS

Hossain M B[1], *Khan S*[2]

[1] *Department of Mathematics & Computing, Australian Centre for Sustainable Catchments University of Southern Queensland, QLD, 4350, Australia and Department of Statistics Biostatistics and Informatics, University of Dhaka, Dhaka 1000, Bangladesh hossainm@usq.edu.au.*

[2] *Department of Mathematics & Computing, Australian Centre for Sustainable Catchments University of Southern Queensland, QLD, 4350, Australia khans@usq.edu.au.*

The odds ratio (OR) is one of the most popular and frequently used indices for measuring the extent of association between exposure and its outcomes in randomised controlled trials (RCTs) and similar studies. Meta-analysis combines data from various independent trials in estimating the overall OR for binary outcomes to make the sample size larger so that the inference based on the combined data is reliable. However, there are situations in which the outcomes are on an ordinal scale with more than two categories. The OR can not directly be used without arbitrarily grouping multiple levels of response into two categories. Moreover, the collapsing of data may cause a loss of valuable information and efficiency. In this study, the generalised odds ratio (GOR) is proposed for summarising the difference between two stochastically ordered distributions of an ordinal categorical variable and used for combining the treatment effect in meta-analysis. A quasi-empirical Bayes method is developed for RCTs using GOR under an independent multinomial sampling procedure. This method is useful for identifying the extreme trials and hence improving the meta-analysis with heterogeneous trials. The estimated confidence intervals for the risk measures are much shorter under the new method than that of others.

References

Agresti, A. (1990) Categorical Data Analysis. Wiley, New York.
Saleh, A. K. Md. Ehsanes, Hassanein, K.M., Hassanein, R.S., Kim, H.M. (2006): Quasi-empirical Bayes methodology for improving meta-analysis. Journal of Biopharmaceutical Statistics 16: 77-90.

Md Belal Hossain *is a PhD candidate at the University of Southern Queensland, QLD, 4350, Australia and an Assistant Professor (on leave) from the Department of Statistics, Biostatistics and Informatics, University of Dhaka, Dhaka 1000, Bangladesh. His main area of interest is medical statistics. Belal is now working on improved meta-analysis for ordinal data. In addition, he has been engaged in research with the clinicians on gastractomy, immunonutrition and fundoplication meta-analysis.*

BAYESIAN METHODS FOR MONITORING CLINICAL INDICATORS

Peter Howley[1] *, Stephen Hancock*[2]

[1]*The University of Newcastle*
c/- Room v123, Mathematics Building, The University of Newcastle, Callaghan, NSW, Aust. 2308
Peter.Howley@newcastle.edu.au

[2]*Health Services Research Group, The University of Newcastle*
Room 365, David Maddison Building, Cnr of King and Watt Streets, Newcastle, NSW, Aust., 2300
Stephen.Hancock@newcastle.edu.au

Bayesian hierarchical models are an integral part of the retrospective analysis and reporting of the Australian Council for Healthcare Standards' (ACHS) clinical indicator (CI) data. The reports provided to individual health care organisations (HCOs), however, could be complemented by tools, such as control charts, to enable HCOs to monitor their performance, rather than relying only on these retrospective reports.

A new control chart for monitoring CI data based upon the beta-binomial posterior predictive (BBPP) distribution was compared with the more commonly used Bernoulli CUSUM chart. Run lengths were simulated based on a factorial design of parameter combinations under 'in-control' and 'out-of-control' conditions. In the more likely situation of the underlying proportion of cases with an event of interest having to be estimated, a parameter space was identified where the BBPP chart was shown to have the desired smaller out-of-control ARL.

The presentation will outline the existing methods for reporting upon the ACHS CIs, the chart based on the BBPP distribution, the method of simulation and the results of the comparison between the chart and the CUSUM chart.

Peter Howley *has held an academic position at the University of Newcastle since 2002 and is currently a Senior Lecturer in the Statistics Discipline. Peter worked with the Health Services Research Group (HSRG) as a research statistician in the mid-late 90s and completed a PhD in Statistics in 2005 focusing on new methods for analysing and reporting upon measures of performance in a clinical setting (clinical indicators). Peter's primary research, in collaboration with the HSRG and Australian Council for Healthcare Standards, continues to focus on Bayesian Hierarchical models, improving the ACHS's CI reporting system and monitoring health care performance.*

WHAT, REALLY, IS THE PROBABILITY OF YOU COMMITTING A CRIME?

Ian Hunt

24/30 Eagle Wharf, London, N1 7EH, UK
ian@tagroup.co.nz

If you are convicted of a crime in New Zealand the Corrections Department attaches a unique probability of recidivism to you. The probability is calculated by the Corrections Department's so-called "objective statistical model"[1]. The Waitangi Tribunal case WAI-1024 upheld the exclusive use of Corrections Department probabilities in sentencing and parole processes[2]. An act of crime is a single-case event driven by human volition - it is difficult to understand or interpret what a probability of such an event could really be. Do recidivism probabilities exist? If so, is there only one for any given case? Are the Corrections Department's estimates reasonable?

I argue that recidivism probabilities do exist because there is a valid and sensible interpretation: calibrated Bayesian probability[3]. This interpretation implies that many different probabilities can be reasonable for any given case – principally because Bayesian probability is by nature a subjective belief. Calibrating beliefs to frequencies constrains subjectivity but a plurality of probabilities remains because there are typically many different frequencies to choose from. Furthermore, valid probabilities should be free to conflict or corroborate one another depending on whose Bayesian belief it is, what data and background information is admitted, and which model is used.

Not all probability estimates are as relevant or reasonable as one another. I argue that, due to methodological flaws, there are serious problems with the Corrections Department probabilities. Flaws include: illegitimate interpretation of single-case probability; the use of only one statistical model; disregard for causal theory and alternative background information; a lack of independent model criticism; and inherent racial profiling. The ruling in WAI-1024 does not address these flaws.

I conclude that if recidivism probability is relevant to a sentencing or parole process then admitting only a single estimate is dangerous and contrary to the adversarial spirit of justice: a range of different probabilities and estimates should be considered.

[1] For an overview of the statistical model (part of what the Corrections Department refers to as its "objective business rules") see: http://www.corrections.govt.nz/policy-and-legislation/cpps-operations-manual/volume-1/i.-reports-general/roc_rol.html

[2] The Waitangi Tribunal is a statutory legal entity in NZ. WAI-1024 was a Tribunal case in which the ruling supported the exclusive use of Corrections Department recidivism probabilities in sentencing and parole decisions.
See: http://www.waitangi-tribunal.govt.nz/reports/downloadpdf.asp?ReportID={A9E5DCD5-98ED-4F5E-B194-CA20C753E74C}

[3] A key requirement for a valid probability interpretation is that the probabilities cohere with Kolmogorov's axioms. Also, probability estimates should spring from a well designed chance set-up (see Freedman, Pisani and Purves (1978), "Statistics", PartVII for a classic introduction). Also, reasonable probability beliefs should be "calibrated" to, or constrained by, empirical frequencies.

Ian Hunt *is a an independent statistical consultant with no affiliations (other than a CStat from the RSS). He holds 3 postgraduate degrees (statistics (Otago, NZ), philosophy of science (LSE, London) and finance (City, London)). He worked as a statistical advisor for the claimant in the Waitangi Tribunal case WAI-1024.*

MODEL SELECTION WHEN FITTING A MIXTURE MODEL TO THREE-MODE THREE-WAY DATA

Lynette A. Hunt [1], *Kaye. E. Basford* [2]

[1] *University of Waikato, Hamilton, New Zealand*
lah@stats.waikato.ac.nz

[2] *University of Queensland, Brisbane, Australia*
k.e.basford@uq.edu.au

This paper investigates the behaviour of some commonly used model selection criteria when using the finite mixture model to cluster three way data containing mixed categorical and continuous attributes. We illustrate the performance of these criteria in selecting both the number of components in the model and the form of the correlation structure amongst the attributes when fitting a mixture model to classical three way data sets.

References

Basford, K.E., and McLachlan, G.J. (1985). The Mixture method of clustering applied to three-way data. J. Classification, 2, 109-125.
Hunt, L.A. and Basford, K.E. (2001). Fitting a mixture model to three-mode three-way data with missing information. J. Classification, 18, 209-226.

Lyn Hunt *currently works as a senior lecturer at the University of Waikato in Hamilton. Her interests are in the clustering of data using mixture models and coping with missing data.*

ASYMPTOTIC COVARIANCE AND OUTLIER DETECTION IN A LINEAR FUNCTIONAL RELATIONSHIP MODEL FOR CIRCULAR VARIABLES

Abdul Ghapor Hussin

Centre for Foundation Studies in Science
University of Malaya, 50603 Kuala Lumpur, Malaysia
ghapor@um.edu.my

This paper discusses the asymptotic covariance and outlier detection procedure in a linear functional relationship model for an extended circular model proposed by Caires and Wyatt (2003). We derive the asymptotic covariance matrix of the model via Fisher information and use the results to detect outliers in the model. Consequently, an outlier detection procedure is developed based on the COV RATIO statistics which have been widely used for similar purposes in ordinary linear regression. We show that the above procedure performs well in detecting outliers via simulation. As an illustration, the procedure is applied to the real data of the wind direction which have been measured by two different instruments.

References

Belsley, D. A., Kuh, E. and Welsch, R. E. (1980). Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. New York: John Wiley & Sons.
Caries, S. and Wyatt, L. R. (2003). A linear functional relationship model for circular data with an application to the assessment of ocean wave measurements. Journal of Agriculture, Biological and Environmental Statistics. 8(2), 153 – 169.
Chatterjee, S., Hadi, A. S. and Price, B. (2000). Regression Analysis by Example. New York: 3rd ed., John Wiley & Sons.

A. G. Hussin *has been teaching Statistics and Mathematics at the Centre for Foundation Studies in Science, University of Malaya, Malaysia for more than 10 years after graduating with PhD from Sheffield University, United Kingdom. My area of research includes Measurements Error models and Directional Statistics. I'm very involved in research activities as well as presenting my work at many seminars and conferences around the world.*

CORRELATION ESTIMATION WITH BIVARIATE CENSORED DATA VIA ALTERNATING REGRESSIONS

Ian James

Centre for Clinical Immunology & Biomedical Statistics, Murdoch University
South Street, Murdoch WA 6150
I.James@murdoch.edu.au

Estimation of correlation when both variables are potentially right censored is confounded by identifiability issues and typically requires assumptions about the nature of the association. Semi-parametric approaches have included the use of copulas, frailties or mixed models. Given that correlation is most meaningful within the context of linear association we consider iterative processes via alternating or 'criss-cross' censored linear regressions to estimate the bivariate correlation, under the assumption that each variable regresses linearly on the other. Two approaches are considered. In the first, value fragments corresponding to re-distributed censored responses from one regression are imputed as weighted explanatory variables in the alternative and the process iterated. Correlation is estimated via the two sets of slope parameters. The second replaces the weighted values by data augmentation. The efficacies of the approaches are compared and illustrated via simulations. Our initial results suggest that both methods compensate well for the censoring even with a significant proportion of cases having both variables censored, with the data augmentation approach slightly less biased. Additional (uncensored) covariates can be readily incorporated. We demonstrate the method via analysis of correlated immunological and virological measures on HIV-1 positive patients from the WAHIV Cohort, which may be incomplete for a variety of reasons.

Ian James *is Professor and Director (Biostatistics) at the Centre for Clinical Immunology and Biomedical Statistics, a joint Murdoch University/Royal Perth Hospital research centre within the Institute for Immunology and Infectious Diseases. His research interests encompass general applied statistics/biostatistics methodology with collaborative interests in the areas of HIV/Hepatitiis, multiple sclerosis and Type 1 Diabetes, particularly in host-viral interactions, pharmacogenetics and genetic association studies.*

MEAN-REVERSION IN INTERNATIONAL REAL INTEREST RATES

Jae H. Kim[1], Philip Inyeob Ji[2]

[1]School of Economics and Finance
La Trobe University, Bundoora, VIC 3086, Australia

[2]Department of Accounting and Finance
Monash University, Clayton, VIC 3800, Australia
Philip.ji@buseco.monash.edu.au

This article examines the mean-reversion property of real interest rates. Many past studies based on univariate methods have reported puzzling outcomes of mean-aversion. We employ panel unit root tests and obtain point and interval estimates of the half-life. More specifically, we use the panel unit root (IPS) test developed by Im et al. (2003) and the inverse normal test of Choi (2001) to test whether a set of real interest rates are stationary. We consider the sub-sampling versions of these tests proposed by Choi and Chue (2007). We conduct half-life estimations based on the bias-corrected bootstrap procedure recently proposed by Kim et al. (2007), adopting the highest density region (HDR) method of Hyndman (1996). This provides a more sensible way of obtaining point and interval estimates for half-life, given the atypical distributional properties of the half-life estimator. Kim et al. (2007) provided Monte Carlo evidence that their bias-corrected bootstrap HDR method performs much better than the conventional methods in small samples. A problem of the conventional methods, including the grid bootstrap that Rapach and Wohar (2004) used, is that their bias-correction procedure can push the model parameter estimates to non-stationarity, even though the underlying model is stationary. As demonstrated in Kim et al. (2007), this will give a half-life estimate of infinity for a stationary mean-reverting time series. Our evidence suggests that, in both industrialized and East Asian emerging capital markets, real interest rates are mean-reverting. We also find that, in less reformed financial markets, real interest rates tend to revert to the mean more quickly than in developed markets.

References

Hyndman, R.J. (1996) “Computing and graphing highest density regions”, *American Statistician*, 50, 120-126.

Kim, J. H., Silvapulle, P. and Hyndman, R, 2007, “Half-Life Estimation based on the Bias-Corrected Bootstrap: A Highest Density Region Approach”, *Computational Statistics and Data Analysis*, 51, 7, 3418-3432.

Philip Inyeob Ji *currently works as a lecturer at the Dept of Accounting and Finance, Monash University. His area of interest is Applied Financial Econometrics.*

CLASSIFICATION OF SYNOPTIC WEATHER TYPES OVER EAST AUSTRALIA USING TWO DIFFERENT OBJECTIVE PROCEDURES

Ningbo Jiang[1], Kehui Luo[2], Paul Beggs[3], Wen Zhou[4]

[1]New South Wales Department of Education and Training
Level 9, 66-72 Rickard Road, Bankstown, NSW 2200
ningbo.jiang@det.nsw.edu.au

[2] Macquarie University
Department of Statistics, Faculty of Science, MACQUARIE UNIVERSITY, NSW 2109
kehui.luo@mq.edu.au

[3]Macquarie University
Department of Environment and Geography, Faculty of Science, MACQUARIE UNIVERSITY, NSW 2109
paul.beggs@mq.edu.au

[4]City University of Hong Kong
School of Energy and Environment, 2/F Harbour View 2, 16 Science Park East Avenue, Hong Kong Science Park, Shatin, N.T., Hong Kong, China
wenzhou@cityu.edu.hk

Determining the dominant, recurring categories of atmospheric circulations is important for understanding the regional and larger-sale climate variability and change and other environmental phenomena. This paper discusses classification of synoptic weather types for the eastern Australian region using two objective procedures: a classic 2-stage procedure (CP1) consisting of obliquely rotated T-mode principle component analysis followed by convergent K-means clustering in comparison with a 2-phase procedure (CP2) based on the batch self-organising map training algorithm. A set of four classifications was obtained for the 52-year NCAR/NCEP dataset from two procedures, one (base classification) from CP1, and three from CP2. These classifications were comparatively examined in such aspects as grouping quality (compactness, separation and topological ordering property), mean type maps, average type frequencies, type lifetime and transitions, variability in the frequency of each synoptic weather type on the seasonal, interannual and decadal scales, and in relation to the Southern Oscillation Index (SOI). The results show that four classifications were inter-confirmative and captured a similar set of major synoptic weather types for the eastern Australian region, each having counterparts in the previous analyses and conforming well to local synoptic experience. In particular, the analysis demonstrates that CP2 has a two-fold utility for map classification purposes. If setting the final neighborhood kernel width ≥1, CP2 performs data projection and provides a flexible means for visualizing the broad distribution of the daily weather patterns in the dataset, as also demonstrated in a few previous applications. By setting the final neighborhood kernel width to zero, this procedure functions as a clustering tool that produces results equivalent to those obtained from CP1. Previous applications of the SOM in synoptic climatology were based on the sequential training algorithm. The findings from this analysis thus indicate a new way of using the SOM for classification of weather maps.

References

Jiang, N. (2010). A new objective procedure for classifying New Zealand synoptic weather types during 1958-2008. International Journal of Climatology doi: 10.1002/joc.2126.
Kohonen, T. (2001). Self-Organizing Maps. Berlin: Springer-Verlag.
Ripley, B.D. (1996). Pattern Recognition and Neural Networks. London: Cambridge University Press.

Ningbo Jiang *currently works as a data analyst with the Educational Measurement and School Accountability Directorate, New South Wales Department of Education and Training. His area of interest is the application of statistical methods in different disciplines including social and physical sciences. Ningbo has been involved in studies of topics such as climate change and variability, air pollution, urban heat island effects and intelligence measurement.*

THE EFFECT OF A PRELIMINARY TEST OF HOMOGENEITY OF STRATUM-SPECIFIC ODDS RATIOS ON CONFIDENCE INTERVALS FOR THESE RATIOS

Paul Kabaila[1], *Dilshani Tissera*[2]

[1] *Department of Mathematics and Statistics, La Trobe University, Victoria 3086*
P.Kabaila@latrobe.edu.au

[2] *Department of Mathematics and Statistics, La Trobe University, Victoria 3086*
dstissera@students.latrobe.edu.au

Consider a case-control study in which the aim is to assess the effect of a factor on disease occurrence. We suppose that this factor is dichotomous. Also suppose that the data consists of k strata, with a 2×2 table for each stratum. A commonly-proposed procedure for the analysis of this type of data is the following (see e.g. Section 13.5 Rosner, 2006). We carry out a preliminary test of homogeneity of the stratum-specific odds ratios. If the null hypothesis of homogeneity is accepted, then inference about the stratum-specific odds ratios proceeds on the assumption that these odds ratios are equal. If, on the other hand, this hypothesis is rejected then inference about the stratum-specific odds ratios is carried out without assuming that these odds ratios are necessarily equal. We examine the effect of this procedure on confidence intervals constructed for the stratum-specific odds ratios. The literature on the effect of preliminary model selection on confidence regions begins with the very important work of Freeman (1989) on the effect of a preliminary test of no differential carryover in a two-treatment two-period crossover trial on the confidence interval for the difference of treatment effects. This literature is reviewed by Kabaila (2009). We find that the preliminary test of homogeneity of the stratum-specific odds ratios has a harmful effect on the coverage probabilities of the confidence intervals for these odds ratios.

References

Freeman, P.R. (1989). The performance of the tow-stage analysis of two-treatment, two-period crossover trials. Statistics in Medicine, 8, 1421-1432.

Kabaila, P. (2009). The coverage properties of confidence regions after model selection. International Statistical Review, 77, 405-414.

Rosner, B. (2006). Fundamentals of Biostatistics, 6th edition. Pacific Grove, CA: Thomson.

Paul Kabaila *holds a Reader in Statistics position in the Department of Mathematics and Statistics at La Trobe University in Melbourne, Australia. He has 73 publications in international refereed journals and is an elected member of the International Statistical Institute. His research interests include: (a) frequentist confidence intervals and prediction intervals that utilize uncertain prior information, (b) the effect of preliminary model selection on confidence intervals and prediction intervals, (c) exact confidence intervals from count data and (d) time series prediction intervals that account for parameter estimation errors.*

TESTING THE SLOPE PARAMETER FOR THE LOGISTIC MODEL IN THE CASE OF HIGH OR LOW SUCCESS PROBABILITIES

Toshinari Kamakura[1], *Masayuki Ohkura*[2]

[1]*Chuo University*
1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan
kamakura@indsys.chuo-u.ac.jp

[2]*Graduate school of Chuo University*
1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan
ohkura@indsys.chuo-u.ac.jp

The logistic model is widely used in the field of medical and industrial applications for detecting differences between populations. The slope parameter may sometimes be very important for discriminating two populations from the viewpoint of statistical testing. However, in the case when the probability of occurrence of the event is very high or low, the phenomenon of separation or monotone likelihood is observed in the fitting process of a logistic model and the Wald test sometimes gives rise to very conservative results. Unfortunately, the R function lm (linear model) presents no message for quasi-separation of the model. In this article we investigate separation or quasi-separation for dataset and calculate the probability of separation or quasi-separation as a function of event occurrence probability.

We propose a new method based on bootstrap resampling techniques and compare the true p-values for the likelihood ratio test, Wald test, and other testing methods. Simulations studies illuminate that our new method keeps nominal p-values even for the high or low event probabilities.

References

Albert, A. and Anderson. J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models, Biometrika, 71, 1-10.

Hauck, Jr., W. W. and Donner, A. (1977). Wald's test as applied to hypotheses in logit analysis, JASA, 72, 851-853

Heinze, G. and Schemper, M. (2002). A solution to the problem of separation in logistic regression, Statist. Med. , 21, 2409–2419.

Toshinari Kamakura *currently works as a Professor of Department of Industrial & Engineering in Chuo University in Tokyo, Japan. His area of interest is survival analysis or statistical reliability analysis in the fields of medicine and industry. He is also interested in machine learning for human motion analysis based on the observations from acceleration sensors and motion captures.*

A MODEL-BASED ESTIMATE OF ANTARCTIC MINKE WHALE ABUNDANCE IN PACK-ICE IN EAST ANTARCTICA

Natalie Kelly[1,2], David Peel[1,2], Mark Bravington[1,2], Nick Gales[2]

[1]Wealth from Oceans National Research Flagship and CSIRO Mathematics, Informatics & Statistics
Castray Esplanade, Hobart ,TAS, 7001, Australia
Natalie.Kelly@csiro.au

[2]Australian Marine Mammal Centre
203 Channel Highway, Kingston, TAS, 7050, Australia
Nick.Gales@aad.gov.au

According to boat-based assessments of Antarctic minke whales (Balaenoptera bonaerensis) undertaken by the International Whaling Commission, there has been an apparent decline in their total (circumpolar) abundance in the decade between the late 1980s and late 1990s. One compelling hypothesis to explain this decline is changes in the position of the pack-ice edge, in concert with changes in the relative proportion of minke whales in the ice, may have decreased the number of animals available to be counted by the research vessels, which operate outside of the pack-ice zone. In order to study minke whale abundance and distribution in pack-ice zones, the Australian Government undertook two aerial surveys in East Antarctica during the 2008/09 and 2009/10 austral summers, respectively. During these line-transect surveys, minke whale observations were made across a spectrum of pack-ice habitats over a 20º slice of longitude adjacent to the Australian Antarctic Territory. A spatial line-transect approach (model-based; Hedley and Buckland 2004) was selected to estimate minke whale abundance across these pack-ice habitats. This method uses quantified relationships between animal density and spatial variables that describe processes that might influence the distribution of an animal in its environment. Animal abundance is then estimated by integrating (numerically) beneath a defined section of the fitted density surface. In the Antarctic minke whale abundance example, abundance is estimated as the product of school encounter rate and mean school size, both of which are modelled across space within a Generalized Additive Model framework. In addition to generally spatial variables, pack-ice habitat was characterised using an amalgam of ice concentration data derived from in situ digital photography and large-scale satellite imagery. It is hoped these results will help towards testing the hypothesis that minke whales have moved into pack-ice zones over the last two decades.

References

Hedley, S. L. and Buckland, S. T. (2004). Spatial models for line transect sampling. Journal of Agricultural Biological and Environmental Statistics 9: 181-199.

Natalie Kelly *currently works as a statistician with CSIRO Mathematics, Informatics and Statistics (CMIS) and with the Australian Marine Mammal Centre, in Hobart; the primary focus of her work is to study the distribution and abundance of animals in Antarctic and sub-Antarctic ecosystems. Over the last three years, Natalie has led a project to study minke whales in pack-ice in East Antarctica. Other research areas of interest include: model-based survey design; and developing systems for field-work (survey) management.*

ADAPTIVE DESIGN IN CLINICAL TRIALS

Patrick Kelly

Sydney School of Public Health, University of Sydney
Edward Ford Building, University of Sydney, Camperdown NSW 2006
p.kelly@sydney.edu.au

Adaptive design is currently a topic of much interest in the pharmaceutical industry. The key feature of these types of designs is that adaptations or changes can be made during a trial, but without the type I error rate becoming inflated. Many types of adaptations are possible, such as the number of interim analyses, sample size re-estimation, dropping of treatment arms and changing endpoints. This talk will explain the underlying statistical methodologies to adaptive designs, in particular, combining evidence from different stages and adjusting for multiple testing. Some of the practical issues to designing these types of trials will be highlighted.

Patrick Kelly *is currently a Senior Lecturer of Biostatistics at the School of Public Health, The University of Sydney. He took up this post in 2007. Prior to this he began in 2000 as a Research Fellow, and then later as a Senior Research Fellow, at the Medical and Pharmaceutical Statistics Research Unit, The University of Reading, UK. It was during his time at Reading he began to investigate and develop statistical methodology for clinical trials. His main research interests to date are survival analysis (especially correlated survival data), pharmacogenomics, adaptive group sequential methods for treatment selection, and more recently clustered randomized controlled trials.*

WAVELET-BASED RESAMPLING OF POINT PROCESSES

Richard D Kenderdine

University of Wollongong, Wollongong, NSW, 2522
richardk@uow.edu.au

Traditional resampling methods require data that are independent. The wavelet transform has been shown to reduce or eliminate dependencies between data. By adopting the philosophy of second-generation wavelets (wavelet lifting), it is possible to resample both one- and two-dimensional point processes and produce surrogates that exhibit characteristics similar to the data. These processes can be homogeneous, non-homogeneous or clustered. Furthermore, in the two-dimensional case, surrogates of marked processes with discrete or continuous marks can be obtained. This talk will briefly describe how the combination of Delaunay triangulation and wavelet lifting methods can yield surrogates of two-dimensional point processes. An example of a clustered point pattern will be used to illustrate the procedure.

References

Davison, A.C., and Hinkley, D.V. (1997). Bootstrap Methods and Their Application. Cambridge University Press.

Illian, J., Penttinen, A., Stoyan, H., and Stoyan, D. (2008). Statistical Analysis and Modelling of Spatial Point patterns}. John Wiley and Sons.

Sweldens, W., and Schroder, P. (1996). Building your own wavelets at home. http://citeseer.ist.psu.edu/old/sweldens96wavelets.html

Richard Kenderdine *has spent the last four years as a part-time PhD student at the University of Wollongong. He has been a sessional tutor with the university since 2002 and a self-employed maths tutor since 1995.*

REGRESSION ANALYSIS USING LONGITUDINALLY LINKED DATA

Gunky Kim[1], Ray Chambers[2]

[1]Centre for Statistical and Survey Methodology
University of Wollongong, Wollongong NSW 2522
gkim@uow.edu.au

[2]Centre for Statistical and Survey Methodology
University of Wollongong, Wollongong NSW 2522
ray@uow.edu.au

Probabilistic data linkage is an attractive data collection option when direct measurement is impossible or extremely costly. One important application is where different data sets relating to the same individuals at different points in time are linked to provide a 'synthetic' longitudinal data record for each individual. However, if the same unique identifier is not available in each of the linked data sets, there is always the possibility that linkage errors in the merged data could lead to such a longitudinal record being actually made up of data items from different individuals. This in turn could lead to bias and loss of efficiency in regression modelling using the linked data. Recent results on unbiased regression inference using longitudinally linked data in the presence of probabilistic linkage errors are described in this paper. They build on the inference framework described in Chambers (2009), and focus on the situation where the linked data are obtained by linking three separate data sources via two possibly dependent linkage operations. For example, these could represent different registers for the same population at different points in time or they could correspond to the situation where a survey sample is linked to two separate registers, one contemporaneous with the survey and the other containing historical information. In the first case one needs to adjust regression modelling to account for linkage errors as well as errors arising from incomplete linkage, while in the second case there is also the important issue of accounting for the impact of the complex survey design when using the linked survey data in regression modelling. Both these scenarios are considered here, and simulation-based results illustrating the gains from taking account of possible linkage errors in regression modelling using probabilistically linked longitudinal data are presented.

References

Chambers, R. (2009). Regression analysis of probability-linked data. Statisphere, Official Statistics Research Series, Volume 4.

Gunky Kim *is currently working as a research fellow in the Center for Statistical and Survey Methodology at University of Wollongong. His current research interest is in the regression analysis of probability-linked data. His other research interests are the semiparametric copula estimation and its application to time series data.*

SUPERVISED PREDICTION IN REGRESSION AND CLASSIFICATION BASED ON VARIABLE RANKING AND A SELECT NUMBER OF PRINCIPAL COMPONENTS

Inge Koch[1], Kanta Naito[2]

[1]School of Mathematical Sciences, The University of Adelaide, Australia
inge.koch@adelaide.edu.au

[2]Department of Mathematics, Shimane University, Japan
naito@riko.shimane-u.ac.jp

We propose a new method for predicting multivariate responses in a regression setting, and for classifying high-dimensional data. Principal Component Regression (PCR) is a well-established method for reducing the number of predictor variables in regression and classification. However, PCR does not take into account the responses; and thus the `wrong' variables may be chosen which can lead to poor prediction.

Bair e*t al* (2006) proposed an efficient prediction method based on supervised principal component regression. They compare each predictor variable with their univariate responses, order the variables essentially by the strength of the correlation between variable and response, and then use just the first PC as the derived predictor. Motivated by the fact that a larger number of principal components results in better regression performance, in Koch and Naito (2010) we extend the method of Bair *et al* in several ways:

- a comprehensive variable ranking is combined with a selection of the *best* number of components for PCR, and
- the regression approach is generalised to multivariate responses.

In addition we extend our regression approach to classification. We use the matrix of vector-valued labels which replaces the multivariate regression responses and Fisher's linear discriminant rule, although our approach is not restricted to this rule. The two-class problem is of particular interest, and based on ideas of Jung and Marron (2009), we show the consistency of the direction vector which we use in the actual classification step.

Applications to simulated and real data demonstrate the performance of the new method both in the regression and classification setting. The new approach is particularly suited to high-dimension low sample size problems, and for these we show that our comprehensive ranking results in a smaller number of predictors and smaller errors than the method of Bair et al.

References

Bair, E., Hasie, T., Paul, D. and Tibshirani, R. (2006). Prediction by supervised principal components. J. Amer. Statist. Assoc. 101, 119-137.

Jung, S. and Marron, J. S. (2009). PCA Consistency in High Dimension Low Sample Size Context. The Annals of Statistics 37, 4104-4130.

Koch, I. and Naito, K. (2010). Prediction of multivariate responses with a select number of principal components. Computational Statistics and Data Analysis (to appear).

Inge Koch *is Associate Professor in Statistics at the University of Adelaide. Her research interests are in the following areas*

- *Analysis of Multivariate and High-Dimensional Data*
- *Linear and Nonlinear Dimension Reduction*
- *Applications of the above to bioinformatics, medicine, climate and finance.*
- *Nonparametric smoothing for univariate and multivariate data.*

The proposed paper is part of an ongoing collaboration with K Naito, in which the authors exploit the power of Independent Component Analysis in dimension reduction, regression prediction and classification.

EVALUATING ASIAN EMERGING STOCK MARKETS VOLATILITY USING GARCH MODELS

Chaiwat Kosapattarapim

School of Mathematics and Applied Statistics
University of Wollongong, Wollongong NSW 2522
ck691@uow.edu.au

GARCH model is a popular time series model to forecast volatility of financial returns. It is well known that financial log returns are not normally distributed and usually exhibit high kurtosis. Therefore, in this study the volatility forecasting performance of GARCH models, with six different error distributions, is evaluated using the daily closing price data form Strait Times Index in Singapore (STI), Kuala Luampur composite Index in Malaysia (KLCI) and Thai Set Index in Thailand (SET). The comparison of out-of-sample volatility forecasts using GARCH models is considered with different error distributions (normal, skew normal, student-t, skew student-t, generalized error distribution (GED) and skew GED). The performance of volatility forecasts for different forecast horizons is studied, using the mean square error (MSE) and the mean absolute error (MAE). Results obtained suggest that a GARCH model with non-normal error distribution gives a better out-of-sample performance than a GARCH model with the normal error distribution. Furthermore, the results obtained suggest that a GARCH model with skew error distribution (skew normal, skew student-t and skew generalized error distribution) has more capability to capture volatility of log returns than a GARCH model with non-skew error distribution.

Chaiwat Kosapattarapim *received his B.Sc. in Statistics from Burapha University, Thailand in 1991 and his M.Sc. in Applied Statistics from Chiang Mai University, Thailand in 1995. He became a lecturer in the Department of Mathematics and Statistics, Faculty of Science, Mae Jo University in 2003. He is currently pursuing his PhD in Statistics at University of Wollongong under a scholarship from Thai Royal government.*

ON MAXIMISING THE LIKELIHOOD OF MARKOV TRIALS

Andrey Kostenko

Monash University, Department of Econometrics and Business Statistics
Clayton Campus, Clayton, Wellington Rd, VIC 3800
Andrey.Kostenko@buseco.monash.edu.au

This study originates with Klotz's (1973) and Devore's (1976) attempts to express the maximum likelihood estimator of the (two-dimensional) parameter θ of a two-state simple Markov chain (Markov trials) explicitly in terms of sufficient statistics. What once could only be imagined is now possible. It is the purpose of this study to report exact and (simpler) approximating expressions for the maximum likelihood estimator θ, as obtained from the full likelihood function in terms of two different sets of sufficient statistics. In addition, an alternative (to that available in the literature) proof of the existence and uniqueness of the solution of the likelihood equations is outlined. As with Bernoulli trials, the maximum likelihood analysis of Markov trials can now be approached with no numerical optimisation routines in a simple and straightforward way.

Andrey Kostenko *is an independent consultant on complex systems. He received his first degree from a state university in Russia and, with the help of a prestigious scholarship from the President of the Russian Federation, an MBA from an institution in the UK. Currently, Andrey is a PhD student with the Department of Econometrics and Business Statistics, Monash University, Australia. His current area of research is statistical forecasting for inventory control, focusing on integer-valued demands that occur randomly in time.*

A FAMILY OF ESTIMATORS FOR POPULATION MEAN USING INFORMATION ON AUXILIARY ATTRIBUTE IN STRATIFIED RANDOM SAMPLING

Nursel Koyuncu[1], *Cem Kadilar*[2]

[1]*Hacettepe University, Faculty of Science, Department of Statistics*
06800, Beytepe, Ankara, Turkey
nkoyuncu@hacettepe.edu.tr

[2]*Hacettepe University, Faculty of Science, Department of Statistics*
06800, Beytepe, Ankara, Turkey

H.S. Jhajj, M.K. Sharma, L.K. Grover, (A family of estimators of population mean using information on auxiliary attribute, Pakistan Journal of Statistics 22 (1), 43-50, 2006) J. Shabbir, S. Gupta (On estimating the finite population mean with known population proportion of an auxiliary variable, Pakistan Journal of Statistics 23 (1), 1-9 2007) and A.M. Abd-Elfattah, E.A. El-Sherpieny, S.M. Mohamed, O.F. Abdou (Improvement in estimating the population mean in simple random sampling using information on auxiliary attribute, Applied Mathematics and Computation 215, 4198-4202, 2010) have suggested some families of estimators by using the known population proportion of elements possessing an attribute in simple random sampling and in two phase sampling. In this paper, after adapting some families of estimators to stratified random sampling, we have proposed a family of exponential ratio type estimators which use the information regarding the population proportion possessing a certain attribute. For the proposed family of estimators, the expressions of bias and mean square error (MSE) up to the first order approximations are derived and the optimum case of the proposed family is discussed in theory. Also an empirical study is carried out to show its properties.

References

Koyuncu, N., Kadılar, C., (2010). On Improvement in Estimating Population Mean in Stratified Random Sampling, Journal of Applied Statistics, 37, 6, 999-1013.

Abd-Elfattah A.M., El-Sherpieny E.A., Mohamed S.M., Abdou O.F. (2010). Improvement in estimating the population mean in simple random sampling using information on auxiliary attribute, Applied Mathematics and Computation 215, 4198-4202.

Nursel Koyuncu *currently works as a research assistant at the Statistics Department in Hacettepe University. She also studied for seven months in REACH study at Erasmus MC medical Center as a researcher. Her area is sample surveys and biostatistics. She has some published papers:*

ABOUT SCALE ESTIMATORS FOR CAUCHY DISTRIBUTION

Olena Kravchuk[1], *Phil Pollett*[2]

[1] *School of Land, Crop and Food Sciences, University of Queensland*
Hartley Teakle Bld, St Lucia, UQ, Brisbane, 4072, QLD, Australia
o.kravchuk@uq.edu.au

[2] *School of Mathematics and Physical Sciences, University of Queensland*
Priestly Bld, St Lucia, UQ, Brisbane, 4072, QLD, Australia
pkp@uq.edu.au

The Cauchy distribution is often used in applications which require thick-tailed or ill-behaved error distribution models. We draw here on the relation between the Cauchy and hyperbolic secant distributions to prove that the MLE of the scale parameter of the Cauchy distribution is log-normally distributed. We also demonstrate that, the Hodges-Lehmann estimator is asymptotically 98% efficient for the scale parameter and performs well even on small samples whether the location of the Cauchy distribution is known or not.

Olena Kravchuk *graduated with a MEng (Automotive Control Systems) from the National Technical University, Ukraine in 1995 and completed a PhD in Statistics at the University of Queensland in 2006. Olena is currently a lecturer in Biometrics and a statistical consultant in the School of Land, Crop and Food Sciences, University of Queensland. Her research interest in mathematical statistics is nonparametrics. She also has a wide range of publications as an applied statistician in Agriculture, Food, Animal and Medical Sciences. Olena is a member of IBS and an accredited member of SSAI (AStat).*

QUANTIFYING POLLUTANT LOADS WITH MEASURED UNCERTAINTY FOR THE BURDEKIN CATCHMENT USING THE 'LOADS REGRESSION ESTIMATOR'

Petra Kuhnert[1], Brent Henderson[2], Stephen Lewis[3], Zoe Bainbridge[4] and John Brodie[5]

[1] CSIRO Mathematics, Informatics and Statistics
Waite Campus, Adelaide SA Australia
Petra.Kuhnert@csiro.au

[2] CSIRO Mathematics, Informatics and Statistics
Canberra ACT Australia
Brent.Henderson@csiro.au

[3] Australian Centre for Tropical Freshwater Research
James Cook University, Townsville QLD Australia
stephen.lewis@jcu.edu.au

[4] Australian Centre for Tropical Freshwater Research
James Cook University, Townsville QLD Australia
zoe.bainbridge@jcu.edu.au

[5] Australian Centre for Tropical Freshwater Research
James Cook University, Townsville QLD Australia
jon.brodie@jcu.edu.au

The export of pollutants from coastal catchments has important implications for the health of the Great Barrier Reef (GBR). As a result, there is a strong need to identify appropriate statistical methods for reliably estimating annual pollutants loads (with some measure of uncertainty) based on monitoring data and assessing progress towards defined loads targets.

The Loads Regression Estimator (LRE), that has been developed under Marine & Tropical Science Research Facility funding, uses statistical methodology for estimating pollutant loads with uncertainties. The approach is regression based and incorporates a four step process: (1) method for flow regularisation to correct for sampling bias; (2) generalised additive model to enable prediction of concentration; (3) the load calculated at regular time intervals; and (4) an estimate of the uncertainty in the loads estimate. The statistical model incorporates terms for flow, and other characteristics of flow (e.g. rising or falling limb, hysteresis and exhaustion effects of the system or flow history), in an attempt to mimic some of the hydrological phenomena observed in these complex systems. Estimates of uncertainty incorporate error in the concentration samples in addition to error in the flow measurements. The regression approach is flexible and can be shown to encompass other existing load estimation methods such as the average estimators.

The methodology has been implemented in the R programming language as the LRE package, which we demonstrate using monitoring data captured in the Burdekin. We will outline the main features of the package and highlight a number of aspects that could be potentially explored to improve its use.

***Dr Petra Kuhnert** is a research statistician in CSIRO's division of Mathematics, Informatics and Statistics with seven years experience working on a wide range of high impact environmental problems. Petra's expertise in uncertainty analysis and in the elicitation of expert opinion and Bayesian methods provides a framework in which to consider the scale and complexity of Australia's environmental challenges. Petra has brought valuable insight and leadership to several important multi-disciplinary high profile projects since her commencement in 2005, including the threat of sediment run-off to the Great Barrier Reef (GBR) and methodologies that incorporate expert opinion in ecological models.*

A METHODOLOGY FOR DECOMPOSING AGE, PERIOD AND COHORT EFFECTS USING PSEUDO-PANEL DATA TO STUDY CHILDREN'S SPORTS PARTICIPATION

Anil Kumar[1], Peter Rossiter[2]

[1]Australian Bureau of Statistics
ABS House, 45 Benjamin Way, Belconnen, ACT 2617
anil.kumar@abs.gov.au

[2]Australian Bureau of Statistics
ABS House, 45 Benjamin Way, Belconnen, ACT 2617
peter.rossiter@abs.gov.au

This paper looks at children's participation in organised sporting activities within a 'pseudo-longitudinal' framework using data from repeated cross-sectional ABS surveys (2000, 2003, 2006 and 2009) of Children's Participation in Cultural and Leisure Activities. The paper examines how children's participation rates vary with age, and also looks for trends over time and any evidence of differences between birth-year cohorts. A simple 'age-period-cohort' (APC) accounting model (proposed by Yang et al., 2008) is initially applied to the pooled data and to selected subpopulations of interest to gain insights into the relative importance of these dimensions. Significant age and period effects are found, but no evidence of cohort effects. To obtain further insights into factors influencing children's participation in sport, especially at the individual level, a logistic regression model is fitted to the data, supplementing the age, period and cohort effects with a range of observed socio-demographic characteristics. The specification of this detailed model is guided by the results of the initial analysis, providing an appropriate solution to the identification problem which arises when APC variables are included simultaneously. The results from the logistic regression model confirm significant age and period effects, and also establish that factors such as gender, parents' employment status, country of birth and the relative socioeconomic status of the neighbourhood are strongly associated with children's participation rates. Children who spend more time on television and computers are also found to be less likely to participate in organised sporting activities.

References

Australian Bureau of Statistics (2000, 2003, 2006a, 2009) Children's Participation in Cultural and Leisure Activities, Australia, cat. no. 4901.0. ABS, Canberra Press.

Glenn, N.D. (2003) "Distinguishing Age, Period and Cohort Effects", in J.T. Mortimer and M.J. Shannan (eds), Handbook of the Life Course, pp. 465–476.
Kluwer Academic/Plenum, New York.

Yang, Y.; Schulhofer-Wohl, S.; Fu, W.J. and Land, K.C. (2008) "The Intrinsic Estimator for Age-Period-Cohort Analysis: What It Is and How to Use It".
American Journal of Sociology, 113(6), pp. 1697–1736.

Anil Kumar *currently works as an Analyst at the Analytical Services Branch of the Australian Bureau of Statistics, in Canberra. His area of interest is in data pooling and socio-economic analysis. Anil has been involved in studying labour force participation of mature age workers, children's participation in sports, creation of synthetic datasets etc.*

GRADUATES' USE OF SOFTWARE FOR QUANTITATIVE ANALYSIS IN THE FINANCIAL SERVICES WORKPLACE

Timothy Kyng[1], Leonie Tickle[2]

[1, 2] Department of Actuarial Studies, Macquarie University
Faculty of Business and Economics, Macquarie University, North Ryde, NSW 2109, Australia
[1] timothy.kyng@mq.edu.au
[2] leonie.tickle@mq.edu.au

We investigate the use of software for quantitative analysis in the workplace by recent graduates and compare this to the software used at university for training these same graduates. In particular we focus on graduates working in the financial services and related industries. The financial services industry is a significant employer of graduates from quantitative disciplines such as Actuarial Science, Mathematics, Statistics and Econometrics as well as financial disciplines such as Economics, Finance and Accounting.

We seek the opinions of recent graduates, their employers and academics who were involved in the education of these graduates. We use questionnaires for this purpose.

This study investigates the use of the following types of software in the workplace by recent graduates. In particular we consider:

- The use of spreadsheets and other financial software
- The use of statistical software such as SAS, S Plus, R, SPSS etc
- The use of other mathematical software packages such as Matlab, Mathematica etc
- The use of industry specific software such as software used for reserving calculations in non life insurance
- The type of software skills required by employers of recent graduates,
- The opinions and attitudes of academics who teach in the relevant disciplines regarding the use of spreadsheets, financial software, statistical and other software used in the teaching and learning of quantitative and financial disciplines.

In industrial practice, skills in using software for model development, model fitting and quantitative analysis are essential. University level training may not equip students to solve the practical problems encountered in the workplace. This study investigates the nexus between learning and work in order to modify the university curriculum. This study is conducted by the Department of Actuarial Studies in the Faculty of Business and Economics at Macquarie University, Australia.

References

Forster, P. A. (2004). Assessing technology-based approaches for teaching and learning mathematics. International Journal of Mathematical Education in Science and Technology 37(2):145-164.

Kyng T. and Taylor, P. (2008). Graduates' use of spreadsheet tools in learning and applying financial mathematics. Asian Social Science 4(3):66-77.

Holton, D. (2005). Tertiary mathematics education for 2024. International Journal of Mathematical Education in Science and Technology, 36(2-3):303-13.

Timothy Kyng *teaches across a range of finance and actuarial studies units at postgraduate level. He is the Director of Postgraduate Studies in the Department of Actuarial Studies, and initiated and implemented the introduction of the new Master of Actuarial Practice degree program in 2006.*

MULTI-LEVEL MODELLING OF DATA FROM A COMPLEX HOUSEHOLD SURVEY: APPLICATION TO THE WESTERN AUSTRALIAN ABORIGINAL CHILD HEALTH SURVEY

David Lawrence

Centre for Developmental Health, Curtin University of Technology and Telethon Institute for Child Health Research, GPO Box K881, Perth. WA. 6842
D.Lawrence@curtin.edu.au

A sample is said to have an informative sampling design if the selection probabilities are related to the values of the dependent variable, even after conditioning on the model covariates. Pfeffermann et al. (1998) described a procedure for fitting a two-level hierarchical model for normally distributed outcomes to data collected from a sample survey with an informative design. This framework has been extended to a three-level model, and also to the case of logistic regression for binary outcome variables. Simulation studies show that the method performs well in estimating the model for the underlying population rather than just fitting a model to the sample data. The methods have been implemented in SAS, and were used in the Western Australian Aboriginal Child Health Survey (WAACHS, Zubrick et al., 2004). The WAACHS was a population-based probability sample of 5,300 Aboriginal children aged under 18 years, and their families, drawn from across the state. The first stage of selection was Census Collection District (CD), the second stage was selection of families and the third stage was children within families. As families from CDs with high numbers of Aboriginal families tended to have different characteristics, this stage of sampling was informative for our models. As incorporating measures of the vagaries of CD design into our models was considered unhelpful, we used our extension of the PWIGLS approach to model the survey data. In practice the technique worked well, with models converging more frequently once the sample design has been properly accounted for.

References

Pfeffermann D., Skinner C.J., Holmes D.J., Goldstein H., Rasbash J. (1998). Weighting for unequal selection probabilities in multi-level models. Journal of the Royal Statistical Society, Series B. 60: 23–40.

Zubrick S.R., Lawrence D., Silburn S.R., Blair E., Milroy H., Wilkes T., Eades S., D'Antoine H., Read A., Ishiguchi P., Doyle S. (2004). The Western Australian Aboriginal Child Health Survey: The health of Aboriginal children and young people. Perth: Telethon Institute for Child Health Research.

David Lawrence *is senior statistician in the Centre for Developmental Health, a joint venture between Curtin University of Technology and the Telethon Institute for Child Health Research. His research interests span a number of areas including mental health, smoking, obesity, diet and nutrition, and Aboriginal health. He was a co-author of the Western Australian Aboriginal Child Health Survey, and has expertise in survey methodology, and the use of linked administrative data.*

CONTROL CHARTS FOR ACCIDENT RATES IN THE CONSTRUCTION INDUSTRY

Nick Fisher[1], Alan Lee[2] and Ross Sparks[3]

[1] Valuemetrics Australia
PO Box 1049, North Sydney, NSW 2059 AUSTRALIA
nif@valuemetrics.co.au

[2] Department of Statistics, University of Auckland
Private Bag 92019 Auckland 1142, New Zealand
lee@stat.auckland.ac.nz

[3] CSIRO Mathematics, Informatics and Statistics
Locked Bag 17, North Ryde, N.S.W. 1670, Australia
Ross.Sparks@csiro.au

Monitoring accident rates in industrial enterprises is an important part of an overall strategy to achieve greater safety in the workplace. In particular, it is important to detect rapidly any upward change in accident rates, so that remedial action can be taken. Control charting techniques have been in use for many years to detect changes in defect rates in industrial production and more recently in medicine and public health, for example to detect changes in adverse event rates, hospital admissions and disease incidence (Woodall, 2006, Grigg and Farewell 2004), but do not seem to have been used in accident monitoring. In this paper we describe some control charting techniques that can be used to give timely warning of changes in accident rates. We cover EWMA, cusum and Shewhart charts, with particular emphasis on steady-state average run lengths, adjustment for risk, and non-Poisson counts.

References

Grigg, O, and Farewell, V. (2004). An overview of risk-adjusted charts. Journal of the Royal Statistical Society, Series A, 167, 523-539.
Woodall, W.H. (2006). The use of control charts in health-care and public-health surveillance. Journal of Quality Technology}, 38, 89-134.

Alan Lee *is Professor of Statistics at the University of Auckland, New Zealand. Among other interests, he collaborates extensively with Dr Nick Fisher of Valuemetrics Australia on a variety of statistical projects, including the monitoring of accidents on the Australian construction industry.*

MULTIPLE IMPUTATION IS NOT ALWAYS HELPFUL: A SIMULATION STUDY WITH IMPERFECT IMPUTATION MODELS AND VARYING FRACTIONS OF MISSING DATA

Lee KJ[1] and Carlin JB[2]

[1]Clinical Epidemiology and Biostatistics Unit
Murdock Childrens Research Institute
Royal Childrens Hospital
Flemington Road, Parkville, Melbourne, 3052
katherine.lee@mcri.edu.au

[2]Clinical Epidemiology and Biostatistics Unit
Murdock Childrens Research Institute
Royal Childrens Hospital
Flemington Road, Parkville, Melbourne, 3052
john.carlin@mcri.edu.au

Multiple imputation (MI) is becoming increasingly popular for handling missing data. However, MI is often implemented without considering how many data are missing nor whether potential gains in precision may be offset by bias introduced by a poorly fitting imputation model. We used a simulation study to compare the bias and precision of regression coefficients estimated using MI and complete case analysis, with varying amounts of missing data. 1000 datasets of 1000 observations were created from a synthetic population, with missingness induced in a highly-skewed continuous covariate or in the binary covariate of interest. Imputations were carried out using multivariate normal imputation (Stata 11), with a simple or a zero-skewness shifted log-transformation to adjust for non-normality. Estimates of the regression parameters were compared to the "true values" from the synthetic population.

When missingness was imposed in the continuous covariate, there was less bias and greater precision for the binary covariate under MI compared with complete case analysis. These findings were consistent irrespective of the amount of missing data, with larger gains in precision as the amount of missing data increased. However, large biases and substantial undercoverage were apparent in estimates of the coefficient for the continuous covariate when there were moderate amounts of missing data and the non-normality was not adequately addressed in the imputation model. When the binary covariate of interest had missing data, although all estimates had negligible bias, gains in precision for its coefficient estimate were minimal.

Although MI can be very useful if missingness is in covariates required for adjustment, gains are much less when there are missing data in the variable of interest. Using a normal imputation model for a skewed covariate led to bias in estimates for the effect of that covariate, even with 25% of data missing.

Katherine Lee *currently works as a Biostatistician in the Murdoch Childrens Research Institute at the Royal Children's Hospital in Melbourne where she has been based for over 2 years. Prior to this she had 2 years experience working as a statistician in a clinical trials unit in London after completing her PhD in medical statistics at Cambridge University. Her primary role in her current position is to provide statistical support for clinical trials and other projects around the hospital. Her methodological work has included exploration of clustering in individual randomised trials and flexible random effects models. Her more recent work has focussed on multiple imputation, with a primary focus on practical aspects of this approach including a comparison of methods for multiple imputation and exploring the potential pitfalls of this approach.*

AN EXPONENTIAL WEIGHTED FUZZY TIME SERIES APPROACH FOR FORECASTING TOURIST ARRIVALS

Muhammad Hisyam Lee[1], Suhartono[2], Hossein Javedani[3]

[1]Department of Mathematics, Universiti Teknologi Malaysia
IT Manager, Research Management Centre, Universiti Teknologi Malaysia
81310 Skudai, Johor, Malaysia
mhl@utm.my

[2]Department of Statistics, Institut Teknologi Sepuluh Nopember
Department of Statistics, Kampus ITS, Keputih Sukolilo Surabaya, Indonesia 60111
suhartono@statistika.its.ac.id

[3]Department of Mathematics, Universiti Teknologi Malaysia
81310 Skudai, Johor, Malaysia
h.javedani@gmail.com

Literature reviews show that the most commonly studied fuzzy time series models for the purpose of forecasting are first order. In such approaches, only the first lagged variable is used when constructing the first order fuzzy time series model. Therefore, such approaches fail to analyze accurately trend and seasonal time series which is an important class in time series models. In this paper, an exponential weighted fuzzy time series is proposed in order to analyze trend and seasonal data and data are taken from tourist arrivals series. In addition, a graphical order fuzzy relationship is proposed to identify the best Fuzzy Logical Relationship order of fuzzy time series. A data set about the monthly number of tourist arrivals to Indonesia is selected to illustrate the proposed method and compare the forecasting accuracy with other weighted fuzzy time series and some classical time series models. The results of the comparison in test data show that the proposed method produces more precise forecasted values than those other approaches.

References

Chen, S. M. (1996). Forecasting enrollments based on fuzzy time series. Fuzzy Sets and Systems, 81(3), 311–319.

Cheng, C. H., Chen, T. L., Teoh, H. J., and Chiang, C. H. (2008). Fuzzy time series based on adaptive expectation model for TAIEX forecasting. Expert Systems with Applications, 34(2), 1126–1132.

Yu, H. K. (2005). Weighted fuzzy time-series models for TAIEX forecasting. Physica A: Statistical Mechanics and its Applications, 349, 609–624.

Muhammad H. Lee *received a BS degree with honors and a MS degree from the Universiti Kebangsaan Malaysia in 1991 and 1993. He also received the PhD degree from the Universiti Teknologi Malaysia, Malaysia in 2002. He joined the Universiti Teknologi Malaysia as a lecturer in 1991. He is currently IT Manager of the Research Management Centre and Associate Professor of Statistics within the Department of Mathematics, Universiti Teknologi Malaysia. He served as an editor and also webmaster for Journal of Matematika. His present research interests include fuzzy time series, forecasting methods, performance evaluation in mobile communications systems, wireless networking, mobile computing, and data analysis. Dr. Lee has been a council member of the Malaysian Institute of Statistics fsince 2008, and a life member of the Malaysian Mathematical Society since 1999. He received a SAS Lecturer Excellence Award at the SAS Forum 2008, KLCC, Malaysia.*

SHRINKAGE ESTIMATION OF A UNIVARIATE NORMAL MEAN

Adityanand Guntuboyina[1], *Hannes Leeb*[2]

[1] *Yale University*
24 Hillhouse Avenue, New Haven, CT 06510
adityanand.guntuboyina@yale.edu

[2] *University of Vienna, Universitätsstr*
5/3, 1010 Vienna, Austria
hannes.leeb@univie.ac.at

It is well-known that there is no Stein phenomenon in dimensions one and two; cf. Stein (1956). In a linear regression setting, we study estimation of the mean of a new response given the corresponding new explanatory variables and a training sample. When the explanatory variables are held fixed, an estimator based on the James-Stein estimator performs poorly when compared to the maximum likelihood estimator in terms of worst-case risk. But on average (with respect to the new explanatory variables), this univariate James-Stein-based estimator dominates the maximum likelihood estimator, irrespective of the unknown parameters. We give an explicit finite-sample analysis of this phenomenon and find, in particular, that shrinkage estimation has certain attractive properties, even when the goal is estimation of a univariate normal mean.

References

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a normal distribution Proceedings of the Third Berkeley Symposium on Mathematics and Statistics, 197-206.

Hannes Leeb *is professor of statistics at the University of Vienna.*

CAN MACHINE LEARNING METHODS BE APPLIED FOR SPATIAL PREDICTIONS OF ENVIRONMENTAL PROPERTIES?

Jin Li[1], Andrew D. Heap[2], Anna Potter[3] and James J. Daniell[4]

[1]Geoscience Australia
GPO Box 378, Canberra, ACT 2601, Australia
jin.li@ga.gov.au

[2]Geoscience Australia
GPO Box 378, Canberra, ACT 2601, Australia
andrew.heap@ga.gov.au

[3]Geoscience Australia
GPO Box 378, Canberra, ACT 2601, Australia
anna.potter@ga.gov.au

[4]Geoscience Australia
GPO Box 378, Canberra, ACT 2601, Australia
james.daniell@ga.gov.au

Spatial interpolation methods for generating spatially continuous data from point locations of environmental variables are essential for ecosystem management and biodiversity conservation. They can be classified into three groups (Li and Heap 2008): 1) non-geostatistical methods (e.g., inverse distance weighting), 2) geostatistical methods (e.g., ordinary kriging: OK) and 3) combined methods (e.g. regression kriging). Machine learning methods, like random forest (RF) and support vector machine (SVM), have shown their robustness in data mining fields. However, they have not been applied to the spatial prediction of environmental variables (Li and Heap 2008). Given that none of the existing spatial interpolation methods is superior to the others, several questions remain, namely: 1) could machine learning methods be applied to the spatial prediction of environmental variables; 2) how reliable are their predictions; 3) could the combination of these methods with the existing interpolation methods improve the predictions; and 4) what contributes to their accuracy? To address these questions, we conducted a simulation experiment to compare the predictions of several methods for mud content on the southwest Australian marine margin. We tested a variety of existing spatial interpolation methods, machine learning methods and their combinations. For the machine learning and combined methods, bathymetry, distance-to-coast, seabed slope, latitude and longitude were used as the secondary variables. The accuracy of the methods was assessed using a 10-fold cross validation. In this study, we discuss results derived from this experiment, visually examine the spatial predictions, and compare the results with the findings in the previous publications. The outcomes of this study have both practical and theoretical importance and can be applied to the spatial prediction of a range of environmental variables for informed decision making in environmental management and conservation. This study reveals a new direction in and provides alternative methods for spatial interpolation in environmental sciences.

References

Li, J., and Heap, A.D. (2008). A Review of Spatial Interpolation Methods for Environmental Scientists. Geoscience Australia, Record 2008/23, 137pp.

Jin Li *currently works as a spatial modeller/computational statistician and is a discipline leader of spatial and statistical modelling in the Marine and Coastal Environmental Group at Geoscience Australia, in Canberra. His area of interest is the spatial prediction of environmental properties and ecological modelling in relation to disturbances and environmental in both terrestrial and marine ecosystems. He has over 20 years' research experience in ecological modelling and/or spatial statistics acquired from his research in the Chinese Academy of Science, the University of New England, CSIRO and Geoscience Australia. Jin is an Associate Editor of an international scientific journal, Acta Oecologica.*

MULTILEVEL SURVIVAL ANALYSIS OF MILD TRAUMATIC BRAIN INJURY IN A COHORT OF NONPROFESSIONAL MALE RUGBY PLAYERS

Ling Li[1], Stephane Heritier[2], Mark Stevenson[3], Stephanie Hollis[4]

[1] Centre for Epidemiology and Research, Department of Health, NSW
73 Miller St., North Sydney, NSW 2060
lynnlynnliau@gmail.com

[2, 3, 4] The George Institute for International Health, The University of Sydney, Sydney
PO Box M21, Missenden Road, NSW 2050
sheritier@george.org.au
mstevenson@george.org.au
shollis@george.org.au

Mild traumatic brain injury (mTBI) is an emerging public health issue in high-contact sports. Nonprofessional rugby has a high incidence of mTBI at individual player level (Hollis et al, 2009). The sustainment of mTBI may be related not only to the risk and protective factors at individual player level but also to the explanatory and management factors at the club level. In this study, we examine the importance of rugby clubs in the sustainment of mTBI among nonprofessional male players based on the time to mTBI with censored data.

A cohort of 1958 male rugby players from Sydney, NSW, was recruited and followed over one to three playing seasons and the mTBI cases were recorded. We observe that the mTBI incidence rates (per 1000 player game hours) vary among different clubs. We use the multilevel discrete-time hazard models and piecewise exponential survival models at both individual player level and the club level with the adjustment for the individual and club level covariates.

There is strong evidence to support the existence of the club level variation in the time to sustainment of mTBI after adjustment for covariates at individual player level, such as the body mass index (BMI), the training hours per week and the number of concussions in past 12 months prior study ($P<0.001$). This variation could be partially explained by the level of support at club and the club competition level.

Reference

Hollis, S.J., Stevenson, M.R., McIntosh, A.S., Shores, E.A., Collins, M.W., Taylor, C.B. (2009). Incidence, Risk and Protective Factors of Mild Traumatic Brain Injury in a Cohort of Australian Nonprofessional Male Rugby Players. The American Journal of Sports Medicine 37(12):2328-33.

Ling Li *currently works as a trainee biostatical officer at the NSW Department of Health. Her area of interest is biostatistics, especially in multilevel and longitudinal modelling. Ling completed her PhD in Statistics at Macquarie University.*

A STATISTICAL DOWNSCALING MODEL FOR SOUTHERN AUSTRALIA WINTER RAINFALL

Yun Li[1], Ian Smith[2]

[1] CSIRO Mathematics, Informatics and Statistics, Wembley, Western Australia
Yun.Li@csiro.au

[2]CSIRO Marine and Atmospheric Research, Aspendale, Victoria, Australia
Ian.Smith@csiro.au

A technique for obtaining downscaled rainfall projections from climate model simulations is described. This technique makes use of the close association between mean sea level pressure (MSLP) patterns and rainfall over southern Australia during winter. Principal components of seasonal mean MSLP anomalies are linked to observed rainfall anomalies at regional, grid point and point scales. A maximum of 4 components is sufficient to capture a relatively large fraction of the observed variance in rainfall at most locations. These are used to interpret the MSLP patterns from a single climate model which has been used to simulate both present day and future climate. The resulting downscaled values provide (a) a closer representation of the observed present day rainfall than the raw climate model values and, (b) provide alternative estimates of future changes to rainfall that arise due to changes in mean MSLP. While decreases are simulated for later this century (under a single emissions scenario), the downscaled values, in percentage terms, tend to be less.

References

Li, Y., and Smith, I. (2009). A Statistical Downscaling Model for Southern Australia Winter Rainfall. Journal of Climate, 22, 1142-1158.

Yun Li *is a senior research statistician with CSIRO Mathematics, Informatics and Statistics (CMIS). His area is in statistical modelling of climatology and oceanography problems. Yun has been involved extensively in climate research projects supported by the Australian Government, Western Australian State Government and CSIRO Climate Adaptation Flagship. He has developed and managed a cross-discipline, international project "Research on Rainfall and Climate Change in both China and Australia" funded through the Australia-China Bilateral Climate Change Partnership, run by the Australian Department of Climate Change.*

HABITAT USAGE OF EXTENSIVELY FARMED RED DEER HINDS IN RESPONSE TO ENVIRONMENTAL FACTORS OVER CALVING AND LACTATION

R.P. Littlejohn[1], *G.W. Asher*[2]

[1]*AgResearch*
Invermay Agricultural Centre, Private Bag 50034, Mosgiel, New Zealand
roger.littlejohn@agresearch.co.nz

[2]*AgResearch*
Invermay Agricultural Centre, Private Bag 50034, Mosgiel, New Zealand
geoff.asher@agresearch.co.nz

Global Positioning Systems (GPS) technology was used to determine the positions of farmed red deer hinds during calving and lactation on an extensively managed high-country station. Meteorological data was collected from a nearby weather station. We describe an analysis of the data relating hind behaviour (half-hourly distance travelled, altitude, habitat occupancy) to environmental factors. Hinds showed strong individualisation in their core occupancy areas, with collared hinds occupying disjoint areas with different aspects and within variable vegetation zones. Heavier hinds selected lower, flatter zones of naturalised grass, while smaller hinds tended to select higher altitudinal zones dominated by tussock. During the pre-calving/parturition period there was no evidence of any influence of weather variables on behaviour, indicating that reproductive behaviours, described by a simple hidden Markov model, took precedence over general behavioural patterns. During the subsequent lactation period, there was clear evidence of diurnal patterns of distances travelled and altitudinal occupation that were moderately influenced by weather variables, with associations between altitude and wind speed, and between distance travelled and solar irradiation and temperature.

Roger Littlejohn *has worked for 27 years as a statistician with AgResearch, at Invermay near Dunedin. A lot of his work has been with the deer group, and he is interested in hidden Markov models and time series analysis.*

EVALUATION OF FEATURE-BASED TIME SERIES CLUSTERING

Shen Liu[1], Elizabeth Ann Maharaj[2]

[1]Department of Econometrics and Business Statistics, Monash University, Caulfield
900 Dandenong Rd, Caulfield East, Victoria 3145, Australia
Shen.Liu@buseco.monash.edu.au

[2]Department of Econometrics and Business Statistics, Monash University, Caulfield
900 Dandenong Rd, Caulfield East, Victoria 3145, Australia
Ann.Maharaj@buseco.monash.edu.au

The clustering of time series is of much interest, and of considerable relevance in many fields of study. Various time series features have been proposed in the literature to cluster time series. For example, Caido et al. (2006) considered periodogram features, while Xiong and Yeung (2004) considered autoregressive mixtures. In this paper, we evaluate the performance of the k-means and k-medoids algorithms in clustering time series when using the following features: autocorrelation function (ACF), partial autocorrelation function (PACF), normalized periodogram (NP), log-normalized periodogram (LNP) and the cepstrum (CEP). We also propose a new weighting system and incorporate it when using the five above-mentioned features. We evaluate the performance of these features, both weighted and unweighted, by comparing their cluster similarity measure values in simulation studies based on stationary time series and variance non-stationary time series. We also consider an application in environmental studies. Overall our results show that the newly proposed weighting system leads to improved values of the cluster similarity measures; the time domain features, namely, the ACF and PACF, tend to achieve better performance for stationary time series; the cepstrum, which is a frequency domain feature, tends to achieve better performance for variance non-stationary time series. In addition, the k-medoids algorithm tends to achieve better performance than the k-means algorithm.

References

Caiado, J., Crato, N., & Peña, D. (2006). A periodogram-based metric for time series classification. Computational Statistics & Data Analysis, 50, 2668 – 2684.

Xiong, Y., & Yeung, D.-Y. (2004). Time series clustering with ARMA mixtures. Pattern Recognition, 37, 1675 – 1689.

Shen Liu *is a Ph.D. student in the Department of Econometrics and Business Statistics, Monash University, Australia. He received his Masters degree in Applied Econometrics from Monash University in 2008. As a co-author, he published a paper in the International Journal of Forecasting in 2010 (forthcoming). His current research interests relate to clustering and forecasting methods of time series.*

A SEAMLESS ADAPTIVE DESIGN WITH REGIMEN SELECTION IN THE UNBALANCED RANDOMISATION RATIO FRAMEWORK

Serigne Lo[1], Stephane Heritier[2], and Caroline Morgan[3]

[1]The George Institute for International Health
PO Box M201, Missenden Road, NSW 2050
slo@george.org.au

[2]The George Institute for International Health
PO Box M201, Missenden Road, NSW 2050
sheritier@george.org.au

[3]Cardinal System
91, avenue de la République, 75011 Paris
c.morgan@cardinal-sys.com

Adaptive designs are becoming ever more popular in clinical trials. Of particular interest are seamless designs that allow a combination of phase II and III data, selection of one or several experimental arms without inflating the type I error rate. In this setting, statisticians face the challenging task of determining the appropriate sample size and sorting out the methodological aspects created by this added flexibility. A 5-arm study in pediatric dermatology coming from our consultancy activity is used as background. Four different doses of an experimental drug have to be compared to a placebo. At the interim analysis one (two) of four doses is (are) selected and carried over to the second stage. The objective is to demonstrate that the selected dose(s) is (are) superior to the placebo at the final analysis. Additional complexity arose in this trial as placebo patients are difficult to enrol due to growing anecdotal evidence in favour of the experimental drug. The extension of the theory to the unbalanced case (k:1 with k>1) is examined. We show that a combination test with appropriate weights can still preserve the type I error but requires a proper choice of the test statistic in small samples. The problem occurs as the asymptotic distribution of the unpooled z-test for proportions is poor in that case, which ruins the performance of the overall procedure. Monte Carlo simulations show that the combination test based on the pooled z-test is still valid but at the expense of a reduced power. Similar problems are expected for other types of outcomes.

Serigne Lo *currently works at the George Institute for International Health, in Sydney. After his PhD in statistics in June 2006, he worked as a biostatistician consultant, and methodological research fellow in clinical trial. His areas of interests are advanced survival analysis (Competing risk; Semi-Markov process), small sample estimation and inference and adaptive design. Serigne has been involved various number of Australian and international clinical trials.*

HIERARCHICAL MODEL-BASED DESIGN OF SURVEILLANCE IN BIOSECURITY, WITH STRUCTURED ELICITATION OF DESIGN PARAMETERS

Samantha Low-Choy[1],*Sharyn Taylor*[2],*Kerrie Mengersen*[3]

[1] *Cooperative Research Centre for National Plant Biosecurity, node at Discipline of Mathematical Sciences, Faculty of Science & Technology, Queensland University of Technology*
2 George St, Brisbane, Queensland 4001, Australia
s.lowchoy@qut.edu.au

[2] *Plant Health Australia, 5/4 Phipps Close, Deakin, Australian Capital Territory, 2600, Australia,*
staylor@phau.com.au

[3] *Collaborative Centre in Data Analysis, Modelling & Computation, Discipline of Mathematical Sciences, Faculty of Science & Technology, Queensland University of Technology*
2 George St, Brisbane, Queensland 4001, Australia
k.mengersen@qut.edu.au

In the plant biosecurity context, the capacity of surveillance systems for early detection of pests (including diseases) and to support claims of area freedom are key requirements of international trade agreements and productivity. Due to the large extent and complexity of import and supply chains, it is difficult to ensure sufficient coverage in a cost-effective manner, and therefore essential to take a risk-based approach to design surveillance. Several methods for designing surveillance are emerging, from qualitative methods to stochastic scenario trees and designs that trade-off costs with either sensitivity (power to detect) or specificity. In addition the paucity of relevant empirical data means that all methods rely on some degree of expert knowledge, typically to quantify design parameters and/or extrapolate from the literature or data on related pests. Here we present a structured approach to elicitation of design parameters. This helps when formulating statistical distributions to represent expert knowledge together with its uncertainty, a core statistical concern. We have investigated how to embed this expression of expert knowledge as statistical distributions within a hierarchical Bayesian model-based approach to design. In this paper we discuss development of the model-based design approach, and structured elicitation, in the context of substantive case studies on early detection of plant pest species. We consider high priority pests for the Australian grains industry. The introduction and establishment of these pests would have significant impact on the Western Australian grains industry, a substantial contributor to national agricultural productivity.

Samantha Low-Choy *is Senior Research Fellow with the Cooperative Research Centre for National Plant Biosecurity (CRCNPB), and sits with the Discipline of Mathematical Sciences within Queensland University of Technology. Her role involves statistical research and collaborative projects, and aims to build capacity for statistics within CRCNPB. A key concern is design and analysis of surveillance for plant pests and disease, to enable post-harvest integrity of grains, and early detection and area freedom for all crops. Other major concerns are pest risk assessment and design of complex experiments, particularly for understanding resistance to chemical treatments. Structured elicitation of expert knowledge is a core underlying component, since it provides important inputs to design and risk assessment. This work builds on her background in applied statistics and Bayesian modelling, and her specific research interests in species distribution modelling as well as expert elicitation, from design and modelling to software and diagnosing cognitive biases.*

SPLINE SMOOTHING USING ROBUST GENERALIZED CROSS-VALIDATION

Mark A. Lukas[1], Frank R. de Hoog[2], Robert S. Anderssen[3]

[1] Mathematics and Statistics, Murdoch University, South Street, Murdoch WA 6150, Australia, M.Lukas@murdoch.edu.au

[2] CSIRO Mathematics, Informatics and Statistics, GPO Box 664, Canberra ACT 2601, Australia, Frank.deHoog@csiro.au

[3] CSIRO Mathematics, Informatics and Statistics, GPO Box 664, Canberra ACT 2601, Australia, Bob.Anderssen@csiro.au

Generalized cross-validation (GCV) is a popular criterion for the selection of the parameter λ in spline smoothing of noisy data. However, it can be unstable and sometimes leads to severe undersmoothing, especially if the sample size n is small. This shortcoming of GCV led to the development of the robust GCV (RGCV) criterion (Robinson and Moyeed 1989), which uses a combination of the GCV score function and $tr(A^2(\lambda))$, where $A(\lambda)$ is the smoothing matrix, with weighting determined by a robustness parameter $\gamma \in (0,1)$. Although RGCV was first proposed over 20 years ago, there has been little investigation of it until recently. In this talk we will discuss recent work showing that for uncorrelated data, RGCV is a practical and effective parameter selection criterion for any size n. Our development of new $O(n)$ algorithms for the calculation of $tr(A^2(\lambda))$ makes it feasible to compute the RGCV score for large n (Lukas, de Hoog and Anderssen 2010). In the analysis of RGCV (Lukas, de Hoog and Anderssen 2008), we use a geometric approach due to Efron to explain the small-sample stability of RGCV. We also derive expressions for the asymptotic inefficiency for both the prediction error and a stronger Sobolev error (involving derivatives) which show that RGCV performs well for any $\gamma \in [0.2,0.4]$, and better than GCV. We will illustrate the results using simulations with cubic smoothing splines.

References

Lukas, M.A., de Hoog, F.R., and Anderssen, R.S. (2008). Spline smoothing using robust GCV. Report 08-154, CMIS, CSIRO.

Lukas, M.A., de Hoog, F.R., and Anderssen, R.S. (2010). Efficient algorithms for robust GCV spline smoothing. J. Comput. Appl. Math., to appear.

Robinson, T. and Moyeed, R. (1989). Making robust the cross-validatory choice of smoothing parameter in spline smoothing regression, Comm. Statist. Theory Methods, 18, 523-539.

Mark Lukas *is a Senior Lecturer in Mathematics and Statistics at Murdoch University, where he has been employed since 1985. He has taught units on calculus, discrete mathematics, and applied and computational mathematics, and has supervised several Honours and PhD projects. His research interests are regularization methods for inverse and ill-posed problems, non-parametric smoothing of noisy data and computational mathematics in general.*

EFFICIENT EDITING FOR PRICE INDEX COLLECTIONS

Carl Mackin

Australian Bureau of Statistics
Perth Office, GPO Box K881, Perth WA 6842
carl.mackin@abs.gov.au

Significance editing is a technique commonly applied in large scale business surveys to reduce the costs of resolving queries in data received. It works by estimating the impact to key survey outputs of resolving individual edit queries and directing resources to the most serious queries. The technique was introduced by Lawrence and McDavitt (1994) and Farwell (2004) discusses problems encountered in the practical application. Price index collections share many of these practical problems but the complex calculations and high profile outputs have meant resource intensive editing strategies have continued to be implemented. This paper discusses scoring functions appropriate for introducing significance editing to price index collections and presents some results from an evaluation of significance based editing to price indexes compiled by the Australian Bureau of Statistics.

References

Farwell, K. (2004). '1352.0.55.066 Research Paper: The General Application of Significance Editing to Economic Collections (Methodology Advisory Committee), Nov 2004', in Australian Bureau of Statistics Website, accessed 31 May 2010, from http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/1352.0.55.066Nov%202004?OpenDocument

Lawrence, D, and McDavitt, C. (1994). 'Significance Editing in the Australian Survey of Average Weekly Earnings'. Journal of Official Statistics, Vol. 10, No. 4, pp. 437-447.

Carl Mackin *currently works as an Assistant Director with the Methodology Development Unit, Statistical Services Branch, Australian Bureau of Statistics, in Perth. His current area of work concerns methodological support for Price Indexes at the ABS.*

DEMAND PLANNING FOR THE WA POLICE ACADEMY USING MARKOV PROCESSES

Taryn Major[1], John Henstridge[2], Anna Munday[3], Alex Maund[4], John Dickson[5]

[1]Data Analysis Australia
97 Broadway Nedlands WA 6009
taryn@daa.com.au

[2]Data Analysis Australia
97 Broadway Nedlands WA 6009
john@daa.com.au

[3]Data Analysis Australia
97 Broadway Nedlands WA 6009
anna@daa.com.au

[4]Data Analysis Australia
97 Broadway Nedlands WA 6009
alex@daa.com.au

[5]Data Analysis Australia
97 Broadway Nedlands WA 6009
dickson@daa.com.au

Workforce planning and forecasting training demand requirements for an organisation such as the Western Australian Police Academy is complex. The problem is not simply one of looking at typical retirement ages but rather understanding career paths in an organisation with a sophisticated rank structure and promotion by merit. To develop such forecasts, a Markov matrix method based upon modelling of promotion and attrition rates was used, an enhancement of a much simpler model developed in a similar context some years ago.

Binomial generalised linear models were fitted to estimate the transition probabilities, incorporating factors including age, gender and previous relevant work experience to better understand and model promotion and attrition rates.

To incorporate the uncertainty that is associated with any forecasting model, a large number of scenarios were produced, based on various informal assumptions regarding future recruitment, promotion and attrition rates within the organisation. The scenario sets were incorporated into an interactive Microsoft Excel tool, allowing the user to consider a range of forecasts with the potential to define strategic planning options based not just on averages, but also peaks and troughs in recruitment and training demand.

Taryn Major *has been a Consultant Statistician at Data Analysis Australia, Perth, since July 2008, after graduating from Macquarie University with a Master of Applied Statistics. Taryn has worked on numerous projects in a range of application areas since joining Data Analysis Australia. Taryn's main work interests include applied statistics and modelling and she particularly enjoys work with a biostatistics focus.*

COMPARISON OF METHODS FOR FIXED EFFECT META-REGRESSION OF STANDARDIZED DIFFERENCES OF MEANS

Michael J. Malloy[1], Luke A. Prendergast[2], Robert G. Staudte[3]

[1]La Trobe University, Bundoora Victoria 3086 Australia
M.Malloy@latrobe.edu.au

[2]La Trobe University, Bundoora Victoria 3086 Australia
luke.prendergast@latrobe.edu.au

[3]La Trobe University, Bundoora Victoria 3086 Australia
R.Staudte@latrobe.edu.au

Given a number of different studies estimating the same effect size, it is often desired to explain heterogeneity of outcomes using concomitant covariates. For very large sample sizes, effect size estimates are approximately normally distributed and a straightforward application of weighted least squares is appropriate. However, in practice within study sample sizes are sometimes small to moderate. When this occurs, the normality assumption of effect estimates may be violated which can then cast some doubt over findings achieved using weighted least squares. Then one may alternatively variance stabilize the effect size estimates and adopt a generalized linear model. This talk will briefly discuss both methods when the effect sizes are the standardized difference of means. Both methods are compared on an example and via simulations which look at the coverage and width of confidence intervals for the meta-regression coefficient of interest. We then present the results of the simulation studies which show that both methods show considerable differences in performance when within study samples sizes are small to moderate in size.

Michael Malloy *is currently a PhD student with the Department of Mathematics and Statistics, Bundoora Campus, in Victoria. His area of interest is Meta-analysis, specifically in Meta-regression. Michael also tutors/demonstrates in second and third year units, which includes teaching the statistical software package R to second year undergraduate students.*

BAYESIAN ANALYSIS OF LARGE SCALE QUANTITATIVE TRAIT LOCI DATA

Louise Marquart[1], Jonathan Keith[2], Kerrie Mengersen[3]

[1] Queensland Institute of Medical Research & Queensland University of Technology
Queensland Institute of Medical Research, PO Royal Brisbane Hospital, QLD, 4029
Louise.Marquart@qimr.edu.au

[2] School of Mathematical Sciences, Monash University
Melbourne, VIC, 3800
jonathan.keith@monash.edu

[3] School of Mathematical Sciences, Queensland University of Technology
GPO Box 2434, Brisbane, QLD, 4001
k.mengersen@qut.edu.au

Quantitative Trait Loci (QTL) analysis is a branch of statistical genetics that attempts to identify the genomic loci responsible for variation in quantitative traits, based on correlations between marker genotype data and the values of quantitative traits. Since the marker loci are generally assumed to have no functional impact on the quantitative traits, detection of a QTL depends upon there being linkage disequilibrium between the QTL and nearby markers. Traditionally, large-scale QTL analysis has focused on two main types of experiment. The first, known as linkage analysis, is based on data from families of related individuals, and estimates the contributions to the variance of the quantitative trait due to individual loci. The second, known as association analysis, uses data from not-necessarily-related individuals to estimate the contribution to the mean of the quantitative trait due to individual loci. The algorithms and concepts employed in the two types of analysis are quite distinct. There is no reason in principle why the two types of analysis cannot be combined, so that both the mean and the variance of the quantitative traits are modelled, and relationship information is included for any individuals for which it is available.

We will discuss a general model encompassing both linkage and association. In this general model we are able to integrate a range of data types including phenotypic and genotypic data, family trees and covariates. This model has been developed in a Bayesian framework, and takes advantage of the ability of the Bayesian paradigm to integrate multiple data types into a single analysis. The model focuses on medium to large scale data sets, and is compared to existing algorithms and software, including PLINK and LOKI.

References

Heath, SC 1997, 'Markov chain Monte Carlo segregation and linkage analysis for oligogenic models', American Journal of Human Genetics, vol.61, no. 3, pp. 748-60.
Heath, SC, Snow, GL, Thompson, EA, Tseng, C & Wijsman, EM 1997, 'MCMC Segregation and Linkage Analysis', Genetic Epidemiology, vol. 14, no. 6, pp.1011-1016.

Louise Marquart *is currently working as a biostatistician at the Queensland Institute of Medical Research, in Brisbane. She is involved in a range of medical and public health research projects. Louise has just completed her honours project at the Queensland University of Technology, where she has been working with her supervisors Kerrie Mengersen and Jonathan Keith on a Bayesian analysis of large scale QTL data.*

ALWAYS TAKE THE WEATHER (DATA) WITH YOU

Alex Maund[1], *Dr John Henstridge*[2]

[1]*Data Analysis Australia*
97 Broadway Nedlands WA 6009
alex@daa.com.au

[2]*Data Analysis Australia*
97 Broadway Nedlands WA 6009
john@daa.com.au

Weather has an enormous impact on our lives; it influences our moods and behaviours both at a personal level and at a social level. There are the obvious correlations: when it's hot we drink more and use more energy to keep ourselves cool through air conditioning units. But these aren't necessarily linear relationships. And then there's seasonality to consider. And just where do you get the data from in the first place?

Data Analysis Australia has spent many years dealing with the weather and its impact on behaviour, particularly relating to the use of utilities. We have also struggled with issues such as inconsistencies in the weather data sourced from BOM (they have an infrequent habit of moving their weather stations) and how reasonable is it to use a single weather station to represent the weather of a larger geographical area?

A series of applications covering modelling of water and electricity consumption, the use of a single weather index and the Data Analysis Australia approach to dealing with the problems presented in using weather data will be discussed.

Alex Maund *has several years of experience in statistical application and modelling, particularly population planning and analysis. Since joining Data Analysis Australia, Alex has been involved in a number of projects including the management, design and analysis of surveys and statistical modelling of data for clients such as the Department of Defence, Synergy and the Water Corporation. Before moving to Australia in 2007, Alex held statistician positions with the UK Government since graduating in 2002 with Honours in Mathematics from the University of Bath. During his years with the Ministry of Defence and Food Standards Agency, Alex worked on a variety of projects ranging from population models and forecasts for the British Army, to the design and analysis of departmental surveys, to statistical critiques of scientific research commissioned by the Department. Alex also co-developed a sampling methodology to assist the Department of Defence in reconciling segments of their Inventory.*

UTILIZING RETROSPECTIVE MATCHING IN WEIGHTED COX AND KAPLAN-MEIER ANALYSES OF TREATMENT STRATEGIES

Elizabeth McKinnon

Centre for Clinical Immunology & Biomedical Statistics, Murdoch University
South Street, Murdoch WA 6150
E.McKinnon@murdoch.edu.au

Background: Physicians managing chronic diseases such as HIV are typically faced with an array of regimens and treatment strategies from which to choose. When results from randomized trials are limited, the ability to utilize data from observational cohorts can be a useful asset in optimizing these choices. However, to provide valid comparisons between treatment options any analysis of observational data requires accommodation of inherent potential selection biases.

Method: To enhance the contribution of observational data to meaningful assessment of alternative therapy strategies we present a simple approach that reduces biases arising from non-random therapy allocation and also enables ready comparative visualizations. Strata comprising matched cases and controls are derived from a retrospective matching process. At completion of the stratification, individual contributions to analyses are down-weighted according to the respective numbers of cases/controls in their stratum to ensure balanced comparisons. Weighted Cox regression and the construction of weighted Kaplan-Meier curves are readily implemented using standard statistical functions in packages. Furthermore, standard output can be utilized to derive test statistics based directly on the weighted Kaplan-Meier curves, such as area-under-the-curve comparisons. We demonstrate desirable small-sample properties of these methods and discuss their advantages over alternatives which require model fitting to address the issue of selection bias.

Application: Retrospective matching is used to identify comparable sets of patients from the Western Australia HIV cohort in order to examine effects of a treatment-switching strategy implemented to maintain initial viral suppression whilst minimizing possible long-term side effects. Weighted Cox and Kaplan-Meier analyses then focus on time from switching/matching to treatment failure.

Bethy McKinnon *is a Research Fellow at the Centre for Clinical Immunology and Biomedical Statistics, a joint Murdoch University/Royal Perth Hospital research centre within the Institute of Immunology and Infectious Diseases. Recent projects with which she has been involved include genetic association studies of Type 1 Diabetes, HIV pharmacogenetic studies, assessment of HIV treatment-induced tissue pathology, and screening for latent TB in refugee children.*

BAYESIAN TESTING FOR DECREASING PROBABILITY DENSITY FUNCTIONS

Ross McVinish[1],Judith Rousseau[2]

[1] University of Queensland
Brisbane QLD 4072, Australia
r.mcvinish@uq.edu.au

[2] Université Paris-Dauphine
Place du Marchal de Lattre de Tassigny, Paris 75016, France
rousseau@ceremade.dauphine.fr

A common problem in statistics is to determine if the distribution P_* that generated a sample Y^{n} of n independent and identically distributed observations belongs to some set $\mathcal{F}_0$ This problem may be stated formally as a test of hypotheses on P_*;

$$H_0: P_* \in \mathcal{F}_0 \quad \text{against} \quad H_1: P_* \in \mathcal{F}_1 \backslash \mathcal{F}_0,$$

where $\mathcal{F}_1$ is some encompassing set of distributions such as the set of distributions with bounded and continuous densities.

In the Bayesian setting, a natural approach is to specify a prior π on $\mathcal{F}_1$ and evaluate the posterior probability that $P_* \in \mathcal{F}_0$, denoted $\pi(\mathcal{F}_0|\mathrm{Y}^{\mathrm{n}})$. Whether this approach gives a reasonable answer will depend on the specification of the prior. For example, if $\pi(\mathcal{F}_0) = 0$ then $\pi(\mathcal{F}_0|\mathrm{Y}^{\mathrm{n}}) = 0$ for all *n* regardless of whether $P_* \in \mathcal{F}_0$ or $P_* \notin \mathcal{F}_0$. Therefore, we are interested in conditions on the prior that lead to consistency of the posterior probability in the sense that;

- If $P_* \in \mathcal{F}_0$ then $\pi(\mathcal{F}_0|\mathrm{Y}^{\mathrm{n}}) \to 1$, in probability, as $\mathrm{n} \to \infty$.
- If $P_* \notin \mathcal{F}_0$ then $\pi(\mathcal{F}_0|\mathrm{Y}^{\mathrm{n}}) \to 0$, in probability, as $\mathrm{n} \to \infty$.

A number of results have been recently obtained for the case where $\mathcal{F}_0$ is a parametric family of distributions (see McVinish et al. (2009) and the review of Tokdar et al. (2010) for further details).

In this presentation, we consider a more challenging case where $\mathcal{F}_0$ is the set of decreasing densities. We take as our prior on $\mathcal{F}_1$ the random histogram prior (section 5.5 of Ghosh and Ramamoorthi (2003)). The simplicity of this prior facilitates the derivation of sufficient conditions on the prior which guarantee the consistency of the posterior probability.

References

McVinish, R., Rousseau, J. and Mengersen, K. (2009). Bayesian goodness of fit testing with mixtures of triangular distributions. Scandinavian Journal of Statistics, 36, 337-354.
Ghosh, J.K., Ramamoorthi, R.V. (2003). Bayesian nonparametrics. New York: Springer-Verlag.
Tokdar, S. T., Chakrabarti, A. and Ghosh, J. K. (2010). Bayesian non-parametric goodness of fit tests. In Frontiers of Statistical Decision Making and Bayesian Analysis}, Eds: Ming-Hui Chen, Dipak K. Dey, Peter Mueller, Dongchu Sun, and Keying Ye.

Ross McVinish *is currently a research fellow with the Australian Research Council Centre of Excellence for Mathematics and Statistics of Complex systems. His areas of research include estimation for stochastic processes (particularly Markov processes), the limiting behaviour of sequences of Markov chains, and Bayesian nonparametric theory.*

THREE STATISTICAL MODELS OF THE IMPACT OF THE WOOLPUNDA SALT INTERCEPTION SCHEME ON IN-STREAM SALINITY IN THE RIVER MURRAY

A.P. Meissner

Department for Water, Land and Biodiversity Conservation
PO Box 240, Berri, SA 5343
tony.meissner@sa.gov.au

The River Murray as it flows through South Australia acts as a drain for the highly saline regional groundwater that contributes to salinity in the river. The largest inflow of salt, approximately 250 t/d, occurs between Lock 3 and Holder downstream. Forty nine bores, above the river valley, were drilled to the underlying groundwater from 1989 to 1992. Pumping of saline ground-water commenced in 1990 to lower the groundwater gradient to the river thus preventing salty water entering the river. The Woolpunda Salt Interception Scheme (SIS) became fully operational in June 1996. A hydrogeological review of the scheme was carried out in August 2000 (Telfer and Way, 2000) that concluded the impact of the scheme for flows less than 10,000 Ml/d was -46.3 EC units at Morgan, SA.

Daily readings of salinity (EC) and flow are taken at Overland Corner 14 km downstream from Lock 3 and salinity at Holder, 26 km downstream from Overland Corner. Weekly average readings were calculated from the daily values from April 20 1992 until Dec 31 1999. A factor (phase) was derived from the drawdown period from April 1992 to June 1996 and the operating period from July 1996 to December 1999. Three statistical linear mixed effect models were examined: - (1) the EC value at Holder was regressed against the EC and flow at Overland Corner and phase, (2) the difference in EC value between Holder and Overland Corner was regressed against flow and phase, and (3) the two sites were converted to factors (sites) and EC regressed against flow, phase and sites. At flows of 5,000 Ml/d the impact of the SIS was estimated, for the three models, to be 30.2 ± 10.1, 34.6 ± 3.4, and 32.4 ± 2.3 EC units respectively.

References:
Telfer, A and Way, D (1999). Waikerie and Woolpunda Salt Interception Schemes Performance Review Using In Stream Data. Australian Water Environments, Kent Town SA 5067

Tony Meissner *currently works as Principal Scientist - Monitoring in the Resources Monitoring Team, Division of Science, Monitoring and Information, Department of Water, Land and Biodiversity Conservation at Berri South Australia. He has had over 40 years experience in agricultural and water resources research and management. In 2005 he spent a year with the Murray-Darling Basin Commission, Canberra contributing to salinity policy. On his return, he managed the Berri Hydrometric Unit which undertook flow and salinity monitoring of the River Murray. In the past 18 months, Tony, a qualified statistician, has contributed to the analysis of hydrological data particularly along the River Murray in South Australia. Tony is retiring at the end of December 2010 and intends to undertake environmental consultancy and statistical analysis of environmental data.*

SEMI-PARAMETRIC REGRESSION USING WAVELETS: AN APPLICATION TO HOUSE PRICES

Daniel Melser[1], *Spiridon Penev*[2]

[1] *School of Finance and Economics, University of Technology Sydney*
Daniel.Melser@uts.edu.au

[2] *School of Mathematics and Statistics, The University of New South Wales Sydney*
S.Penev@unsw.edu.au

Wavelet smoothing methods for noisy signals are widely used in empirical work. They have proven to have some significant advantages over other smoothing methods in their ability to approximate irregular signals. However, little work has been done on utilizing wavelet smoothing techniques within a more general parametric regression framework. In this paper we seek to develop estimators of semi-parametric regression models which include a standard parametric component as well as a wavelet smoothing function. It is hoped that this will help to broaden the range of applications in which wavelet smoothing methods can be used. We propose to extend our methodology for use in both one and two-dimensional problems. In particular, our approach is likely to be especially useful with regard to the problem of the hedonic estimation of house prices. This is the context in which we make use of this method. We apply our methodology to a large dataset of house prices for Sydney, which includes both structural information as well as a property's location on a spatial grid. Much of the cross-sectional variation is due to differences in dwellings' structural characteristics, however much is also due to location-specific characteristics. We use the wavelet smoother to model this latter influence upon price. Estimated house prices are important in automated valuation methods, which banks use to evaluate how much they should loan on a specific property.

Dr Daniel Melser *is currently working as a Senior Lecturer at UTS http://datasearch.uts.edu.au/business/staff/finance/details.cfm?StaffId=9976 but is also enrolled (part time) as a Masters of Statistics student at UNSW. Dr Spiridon Penev is his supervisor. Daniel Melser has previously completed a PhD in Economics at UNSW (awarded in 2005). His research interests are mainly in the area of economic statistics (CPI, GDP and the national accounts), productivity and effciency, also flexible estimation methods and wavelets.*

USING MODEL CHOICE METHODS FOR MODEL CHOICE IN MIXTURES – A BAYESIAN PERSPECTIVE

Kerrie Mengersen

Queensland University of Technology, Mathematical Sciences
GPO Box 2434, Brisbane 4001
k.mengersen@qut.edu.au

Choosing between competing models is a standard aim in statistical analysis. However, how suitable are these approaches for this purpose? How do they perform in theory and in practice? In this presentation we review some of the common approaches to model choice in a Bayesian framework. We focus on mixture distributions as a concrete context for this review, and present some new methodological and applied results.

This work is collaborative with Professors Christian Robert and Judith Rousseau, and members of the QUT Bayesian Research and Applications Group.

Kerrie Mengersen *holds a Research Chair in Statistics at QUT. Her research interests are in Bayesian modelling and computation, and analysis of complex systems.*

ANSWERING IMPORTANT BUT COMPLICATED QUESTIONS: A STUDY OF THE FACTORS THAT INFLUENCE RESOURCE USE IN AUSTRALIAN CITIES

Denny Meyer[1], *Peter Newton*[2]

[1] *Swinburne University of Technology*
John St, Hawthorn, VIC3122
dmeyer@swin.edu.au

[2] *Swinburne University of Technology*
John St, Hawthorn, VIC3122
pnewton@swin.edu.au

What are the most important determinants of urban resource consumption? This is an important question at a time when current levels of domestic consumption in Australia need to be reduced. Answers to this question will tell us where interventions for winding back consumption should be focused. However, to provide these answers is a challenging task because of the wide range of influencing factors that exist and the multi-faceted nature of resource consumption. This paper uses a variety of measurement and structural models to address this issue. The analysis is based on the results of a survey that was carried out in Melbourne in 2009. The survey was designed to provide accurate measures of water, energy and appliance consumption with more abstract measures of transport and dwelling space consumption, allowing the formation of a total measure for resource consumption. In addition data was collected for the possible determinants of resource consumption in terms of the location, dwelling and, household contexts, individual demographics and behaviours. For each type of consumption general linear models were used to explain the volume of per capita resource consumption in terms of the variables within the five determinant categories. Predictions from these models were then combined in order to evaluate the importance of each determinant category for each type of resource consumption. A final structural model was used to explain the relationship between the determinant categories and total resource consumption, with the household context, in particular household size, playing a pivotal role. This study suggests that it is possible to answer impossible questions if you have the right data and the right statistical tools. Our results show that the determinants of resource consumption vary for the different domains of consumption and that individual attributes are less influential than contextual factors when trying to explain resource consumption.

References

Newton, P. (2006). Australian State of the Environment 2006: Human Settlements Theme paper, Department of Environment and Heritage, Canberra (www.environment.gov.au/soe)

Rose, H. (2003). M3-Simulation Multidisciplinary Simulation of Sustainability Strategies. Systems Analysis Modeling Simulation, 43(9), 1243-1247.

Lenzen, M. Energy and Greenhouse Gas Cost of Living for Australia during 1993/94. Energy, 23(6), 497-516.

Denny Meyer *is a senior lecturer in statistics at the Swinburne University of Technology in Melbourne. She has previously worked at universities in South Africa and New Zealand. She has co-authored two books and has published upward of 50 articles in a variety of refereed journals. She is an applied statistician, working in areas such as management, advertising and social research, specializing in time series analysis, structural equation modelling and data mining.*

CHARACTERISING THE SPREAD OF NEOLITHIC CULTURE ACROSS EUROPE

Steven Miller

Department of Statistics, University of Waikato
Private Bag 3105, Hamilton 3240, New Zealand
smiller@stats.waikato.ac.nz

The Neolithic era, the last major sub-division of the Stone age, began about 11,000 years ago when agriculture was developed in the Middle East. From there, Neolithic culture spread into Europe, with the era ending in Europe about 5000 years ago, upon the widespread adoption of metal tools. The spread of Neolithic culture could have resulted from a migration of people, a diffusion of ideas, or more likely a combination of the two. Residual signals from this transition across Europe might still be found in ancient and potentially modern patterns of genetic and linguistic variation, as well as in the archaeological record. We are attempting to formally model this process by combining evidence from these multiple sources in association with a non-trivial diffusion model across a landscape. This has required the development of a method involving indirect inference to extract information from a series of computer simulations in order to estimate distributions for the diffusion parameters of interest.

Steven Miller *is a lecturer in statistics at the University of Waikato in Hamilton, New Zealand. His interest in the Neolithic transition arose from a research fellowship at Trinity College Dublin, Ireland in 2008-2009. He is particularly interested in how methods developed for characterising ancient continental human migration events might be adapted to suit the relatively more recent colonisation of the Pacific islands, and migrations on much smaller geographical and temporal scales, such as diffusions of recently introduced pest species.*

POISSON LOGLINEAR MODELS FOR POINT PROCESSES: NEGLECTED EXPONENTIAL FAMILY STRUCTURE

Adrian Baddeley[1], Andrew Hardegen[2], Robin K. Milne[3]

[1] CSIRO Mathematics, Informatics and Statistics
Private Bag 5, Wembley WA 6913
Adrian.Baddeley@csiro.au

[2] School of Mathematics and Statistics, M019, University f Western Australia
35 Stirling Highway, Crawley WA 6009
hardegen@maths.uwa.edu.au

[3] School of Mathematics and Statistics, M019 University of Western Australia
35 Stirling Highway, Crawley WA 6009
milne@maths.uwa.edu.au

Poisson loglinear regression models are widely used (e.g. Baddeley et al., 2010) for statistical analysis of point process data. They are parametric models where the natural logarithm of the intensity function of an inhomogeneous Poisson process is linearly related to covariates defined on the base space of the point process. Such models possess an exponential family structure the essence of which, with hindsight, was already apparent in Cox and Lewis (1966). Barndorff-Nielsen and Cox (1994) mentioned the likelihood function for an inhomogeneous Poisson process and gave a prominent place to exponential family theory, yet did not consider loglinear Poisson processes in their own right as examples of exponential families.

Just as a point process can be regarded as a fundamental extension of a collection of jointly distributed random variables (giving the locations of finitely many points in some base space), so the Poisson loglinear regression model for a point process is an extension of the familiar Poisson loglinear regression model for a count response.

The exponential family structure of Poisson loglinear regression models for point processes is reflected in simplicity of structure for likelihood-based inference for such processes. The paper outlines these basic ideas.

Exponential family structure is further exploited in two reductions involving a specified pixel grid. Initially, point process data which is complete, in the sense of having full information concerning all locations, is reduced to corresponding counts for pixels of the grid. Such a reduction is a part, often unavoidable, of many point process applications. Some applications involve a further reduction of such data to presence-absence data for the pixels of the given grid. It is helpful to view these reductions as examples of incomplete data arising from (complete) data which follows an exponential family distribution, and exploit the structure so inherited.

References

Baddeley, A., Berman, M., Fisher, N.I., Hardegen, A., Milne, R.K., Schuhmacher, D. and Shah, R. (2010). Spatial logistic regression and change-of-support in Poisson point processes. Electronic J. Statistics, under revision.

Barndor_-Nielsen, O.E. and Cox, D.R. (1994). Inference and asymptotics. London: Chapman and Hall.

Cox, D.R. and Lewis, P.A.W. (1966). Statistical analysis of series of events. London: Methuen (now Chapman and Hall).

Robin Milne *is a expatriate New Zealander who completed a PhD at the Australian National University in 1971 working with Professors Pat Moran and David Vere-Jones. He has held positions at Victoria University of Wellington NZ, London School of Economics, and the University of Western Australia. He has wide interests in probability and statistics, in particular in point processes.*

GENOMIC ANALYSIS OF A COMPLEX TRAIT WITH SIMULTANEOUS ANALYSIS OF ALL SNP AND GENE EXPRESSION DATA

Allan Motyer[1], *Sue Wilson*[2], *Sally Galbraith*[3]

[1] *Prince of Wales Clinical School*
University of New South Wales
A.Motyer@unsw.edu.au

[2] *Prince of Wales Clinical School and School of Mathematics and Statistics*
University of New South Wales
sue.wilson@unsw.edu.au

[3] *Prince of Wales Clinical School and School of Mathematics and Statistics*
University of New South Wales
sally.galbraith@unsw.edu.au

Genome wide association studies aim to identify genetic causal variants associated with a clinical trait. Typically SNPs (single nucleotide polymorphisms) are tested one at a time to identify genetic variants with the greatest marginal association. For many complex diseases, however, it is more likely that there are multiple causal variants and methods are required that take this into account. We consider a penalised maximum likelihood approach to model selection, based on the method proposed by Hoggart et. al., for the simultaneous analysis of all genetic variants in the case where the number of genetic variants p far exceeds the number of observations n (p >> n). The approach is applied to both SNP and gene expression data from a study into mice obesity (Wang et. al.). Both cross-validation and resample model averaging are used for model assessment.

References

Hoggart, C.J., Whittaker, J.C., De Iorio, M. and Balding, D.J. (2008). Simultaneous Analysis of All SNPs in Genome-Wide and Re-Sequencing Association Studies. PLoS Genetics, 4 (7), e1000130.
Wang, S., Yehya, H., Schadt, E.E., Wang, H., Drake, T.A., and Lusis, A.J. (2006). Genetic and Genomic Analysis of a Fat Mass Trait with Complex Inheritance Reveals Marked Sex Specificity. PLoS Genetics, 2 (2), e15.

Allan Motyer *is a postdoctoral research fellow at the Prince of Wales Clinical School, University of New South Wales. His area of interest is the development and application of novel statistical approaches to analysis of complex genomic data.*

GOODNESS-OF-FIT TEST OF THE MARK DISTRIBUTION IN A POINT PROCESS WITH POSITION DEPENDENT MARKS

T. Mrkvicka[1], *S. Soubeyrand*[2], *J. Chaduf*[2]

[1] *Institute of Mathematics and Biomathematics , Faculty of Science, University of South Bohemia, Branisovská 31, 37005 Ceské Budejovice, Czech Republic*
mrkvicka@prf.jcu.cz

[2] *INRA, UR546 Biostatistique et Processus Spatiaux, F-84914 Avignon, France*

Marked point processes can be used to describe positions and characteristics (height, diameter ...). These characteristics may depend on unobserved spatial fields (soil properties for example). For such point processes the marks are then position dependent. The knowledge of the parametric distribution of marks can be of interest for itself or it can be used for better inference.

Four procedures for testing the hypothesis that the marks of the point process belong to a given family of distributions, where the marks are independent conditionally on an unknown non-stationary parametric field $\theta(x)$, are described. The unknown parametric field is estimated by a kernel estimator and the estimate is included in the procedures. The procedures are compared by a simulation study and they are applied to two real datasets. First, the terminal dates of the Maya sites and second, the heights of the trees felt during storms.

Tomas Mrkvicka *currently works as an assistant professor at the Institute of Mathematics and Biomathematics, University of South Bohemia in the Czech Republic. His area of interest is stochastic geometry and spatial statistics. Mainly he is developing new statistical methods but he also consults with biologists on their statistical problems.*

THE CONVERGENCE OF LEAST SQUARE ESTIMATOR FOR FIRST ORDER GENERALIZED STAR (GSTAR) MODEL

Utriweni Mukhaiyar[1], Udjianna Pasaribu[2], Khreshna Syuhada[3]

[1] Statistics Research Division-Institut Teknologi Bandung (ITB)
Jalan Ganesa 10 Bandung
utriweni@math.itb.ac.id

[2] Statistics Research Division-Institut Teknologi Bandung (ITB)
Jalan Ganesa 10 Bandung,
udjianna@math.itb.ac.id

[3] Statistics Research Division-Institut Teknologi Bandung (ITB)
Jalan Ganesa 10 Bandung,
khreshna@math.itb.ac.id

We consider the properties of the Least Square (LS) estimator for parameters of the first order Generalized Space Time Autoregressive (GSTAR) model. Specifically, we aim at investigating the convergence of these estimators, through its mean squared error, by using Monte Carlo simulation. We take the case when the GSTAR model has normally distributed errors as well as Martingale difference errors. It is found that there should be a minimum number of observations (from the `time' point of view) to guarantee that the estimators are very close to their true parameters.

References

Borovkova, S., Lopuha·· a, H.P., Nurani, B. (2008). Consistency and asymptotic normality of least squares estimators in generalized STAR models. Statistica Neerlandica. 62(4), 482-508.
Box, G.E.P., Jenkins, G.M. (1970). Time Series Analysis: Forecasting & Control. San Fransisco: Holden-Day Inc.
Wei, W.W.S. (2006). Time Series Analysis: Univariate and Multivariate Methods, ed. 2. Boston: Pearson Addison Wesley

Utriweni Mukhaiyar *is a PhD Student at the Institut Teknologi Bandung (ITB), Indonesia, and also works as a Lecturer there. Her research areas include time series, geostatistics and space time series.*

ESTIMATING INFLECTION POINTS OF MOTOR EVOKED POTENTIAL MEASUREMENTS: AN APPLICATION OF MONOTONE SMOOTHING SPLINES

Samuel Müller

University of Sydney, School of Mathematics and Statistics, Australia
samuel.mueller@sydney.edu.au

An application of monotone smoothing splines is presented for the estimation of inflection points of functional monotone data. Neurological data – TST amplitude measurements - was available for 29 healthy subjects. For the estimation of each of the 29 monotone curves the smooth.monotone function in the R-package fda (Ramsay et al., 2007) was used. Its inflection points - the modes of the corresponding derivatives - were of particular neurological interest, i.e. to quantify the contribution of slow-conducting motor tract portions to the TST amplitude. The statistical analysis provided empirical evidence of a bimodal distribution of central conduction times, which might possibly relate to different fibre types within the pyramidal tract. This presentation focuses on the methodological aspects and challenges of selecting similar model parameters for 29 different data sets.

References

Ramsay JO, Wickham H, Graves S. fda (Functional Data Analysis), R package version 1.2.3, 2007; software available at http://www.r-project.org.

Firmin L, Müller S, Rösler KM, 2010, 'A method to measure the distribution of latencies of motor evoked potentials in man', Clinical Neurophysiology, to appear.

Samuel Müller *is Senior Lecturer in Statistics with research interests in statistical model selection, extreme value theory and applied statistics. He graduated from the University of Bern (Switzerland) in 2002 and held positions at the ANU and the University of Western Australia before moving to the University of Sydney in 2008. He regularly publishes in high quality journals of different scientific areas and presents at international conferences.*

GEOSTATISTICAL SIMULATION OF MULTIVARIATE DATA USING MAF AND ACDC METHODS

Ute Mueller

School of Engineering, Edith Cowan University
100 Joondalup Drive, Joondalup, WA, 6027, Australia
u.mueller@ecu.edu.au

Many of the data to be modelled in geostatistics are multivariate, making it necessary to use multivariate estimation or simulation techniques. In particular, the fitting of a suitable covariance model is problematic and so it is common to transform the set of attributes into spatially uncorrelated factors that can be simulated independently. The main method in recent years has been the method of minimum/maximum autocorrelation factors (MAF) (Desbarats and Dimitrakopoulos (2000)), which can be regarded as a spatial principal component analysis. The decorrelation is either based on the coefficient matrices of a two structure linear model of co-regionalisation (LMC) or on a pair of experimental covariance matrices. In both cases it is assumed that the covariance structure of the data can be adequately modelled using a two structure LMC. An extension to spatial decorrelation using experimental semivariogram matrices for a larger set of lag spacings is possible, if an approximate decorrelation is acceptable. There are several algorithms which approximately diagonalise a set of symmetric matrices, one of which is the Alternating Columns-Diagonal Centres (AC-DC) method (Yeredor (2002)). The MAF and AC-DC methods will be illustrated through application to a multivariate data set from a Nickel mine. The resultant factors will be simulated using sequential Gaussian simulation. The extent to which factors obtained from each method are spatially decorrelated will be evaluated along with an assessment of the effect of the transformation method on the simulated attributes.

References

Desbarats, J A and Dimitrakopoulos, R, 2000. Geostatistical Simulation of Regionalized Pore-Size Distributions Using Min/Max Autocorrelation Factors, Mathematical Geology, 32(8):919-942.
Yeredor, A, 2002. Non orthogonal joint diagonalization in the least square sense with application in blind source separation, IEEE Signal Processing, 50(7):645-648.

Ute Mueller *is an Associate Professor in Mathematics at Edith Cowan University. Her research interests include geostatistical simulation algorithms and applications of geostatistical methods to environmental, mining and fisheries data.*

MULTIVARIATE STATISTICS IN TAX ADMINISTRATION

Xin Wang[1], Bhaskaran Nair[2], Michael Slyuzberg[3], Graeme Buckley[4]

New Zealand Inland Revenue
12-22 Hawkestone Street, Wellington 6140
[1] xin.wang@ird.govt.nz
[2] bhaskaran.nair@ird.govt.nz
[3] michael.slyuzberg@ird.govt.nz
[4] graeme.buckley@ird.govt.nz

National Research Unit (NaRU) of New Zealand Inland Revenue (IR) is responsible for development, implementation, monitoring and evaluation of IR business and strategies. NaRU contributes to IR by conducting research, providing advice on analytical practice and methods, and coordinating research efforts. The main objectives of NaRU are to develop robust methodology for statistical analysis, to conduct high quality research with application of sophisticated statistical techniques, and to provide innovative ideas and practice to improve IR business.

Recently NaRU successfully finalised a set of important projects using advanced statistical analysis and data mining techniques. Some of them were:

- Relationships between customers' compliance behaviour and location;
- Customer segmentation focusing on demographics, interactions with IR and compliance behaviour;
- Relationships between macroeconomic fluctuations, business demographics and tax revenue collected by IR;
- Impact of IR debt intervention on customers' compliance;
- Analysis of customers' satisfaction and their perceptions of tax administration.

These employed sophisticated statistical techniques such as logistic regression, classification, segmentation, principal component and factor analysis, and covered the following business and analytical areas:

- risk analysis/management;
- customer intelligence;
- customer relationship management;
- business improvement; and
- market research.

This presentation focuses on the issues encountered with the application of multivariate statistical techniques to "customer segmentation" and "relationship between tax compliance behaviour and location". In regard to customer segmentation, issues of data diversity, feature selection, segmentation techniques and business expectation/requirement are discussed and a business-focused hierarchical scheme, built in a divisive structure and embracing both manual partitioning and the data mining approach, is introduced. For the "relationship between tax compliance behaviour and location", class imbalances in modelling and the problem that poses on the results are discussed.

References

Japkowicz, N. (2000). The class imbalance problem: Significance and strategies. In: Proceedings of the 2000 international conference on artificial intelligence (IC-AI'2000): Special track on inductive learning, Las Vegas, Nevada.

Chawla, N.V. Japkowicz,N and Kotcz,A. (2004) "Editorial: Special issue on learning from imbalanced data sets," SIGKDD Explorations, vol. 6, no. 1, pp. 1–6, 2004.

Hardie, D., Slyuzberg, M., Nair, B., Buckley, G. and O'Connor, M. (2009) Differences in SME Tax Compliance: What Matters? 22nd Annual SEAANZ Conference, 31 Aug 2009, Massey University, Wellington.

Bhaskaran Nair *currently works as Senior Researcher with the New Zealand Inland Revenue. His current area of interest is data mining in tax administration data and customer compliance behavior, besides providing statistical consultancy. Bhaskaran has expertise in the application of statistical techniques to qualitative and quantitative data derived from diverse fields such as education, economics, agriculture and industry, and publication of research reports, papers and documents in both the public and private sectors.*

ANALYSIS OF POINT PROCESSES ON A LINEAR NETWORK

Qi Wei Ang[1],Adrian Baddeley[2],Gopalan Nair[3]

[1] The University of Western Australia
School of Mathematics and Statistics 35 Stirling Highway, Crawley, Perth, Western Australia, 6009
aqw07398@hotmail.com

[2] The University of Western Australia and CSIRO Mathematics, Informatics and Statistics
CSIRO, 65 Brockway Rd, Floreat, WA
adadrian@maths.uwa.edu.au

[3] The University of Western Australia, School of Mathematics and Statistics
35 Stirling Highway, Crawley, Perth, Western Australia, 6009
gopal@maths.uwa.edu.au

The aim of this talk is to present statistical methodologies developed recently to analyse point patterns that occur on a linear network such as traffic accidents on a road network. Applying standard statistical techniques designed to analyse point patterns in two-dimensional space, such as use of the Ripley's K function, does not take into account the constraint that spatial points can only occur on line segments. We propose a modified version of the Ripley's K function, which is developed as a statistic to summarise point patterns on linear networks. The techniques developed will be used to analyse the data giving locations of small spiders that nest across the gaps between bricks in a brick wall.

Gopalan Nair *is an Associate Professor in the School of Mathematics and Statistics at the University of Western Australia in Perth. His areas of interest are general theory of point processes, spatial analysis and queuing theory.*

WHAT LEVEL OF STATISTICAL MODEL SHOULD WE USE IN SMALL AREA ESTIMATION?

[1]Mohammad-Reza Namazi-Rad and [2]David Steel

[1]Centre for Statistical and Survey Methodology
University of Wollongong, NSW 2522, Australia
Mrnr727@uow.edu.au

[2]Centre for Statistical and Survey Methodology
University of Wollongong, NSW 2522, Australia
dsteel@uow.edu.au

If unit-level data are available, small area estimation is usually based on models formulated at the unit level but they are ultimately used to produce estimates at the area level. This paper investigates the circumstances when directly using an area-level model is more effective. Linear mixed models fitted on different levels of data are applied in small area estimation to derive a synthetic estimator and EBLUP. The performance of area-level models is investigated compared with unit-level models when both individual and aggregate data are available. A key aspect is whether there are substantial contextual or area-level effects in the covariates. Ignoring these effects in unit-level working models can cause biased estimates. This is referred to as the ecological fallacy. The mentioned effects can be automatically accounted for in the area-level models. Using synthetic and EBLUP techniques, small area estimates produced based on different levels of linear mixed models are studied in a simulation study.

Reference

Ghosh, M., and Rao, J. N. K. (1994) Small Area Estimation: an Appraisal. Statistical Science. 9, 55-93.

Khoshgooyanfard, A., and Taheri Monazah, M. (2006). A Cost-Effective Strategy for Provincial Unemployment Estimation: A Small Area Approach. Survey Methodology. 32, 105-114.

Longford N. T. (2005). Missing Data and Small Area Estimation. Spring-Verlag.

Mohammad-Reza Namazi-Rad *is a PhD candidate in Applied Statistics studying at the University of Wollongong, Australia. He has a bachelors' degree in Mathematical Statistics and a masters' degree in Social and Economic Statistics. For his PhD he is working on a major research program within the Centre for Statistical and Survey Methodology (CSSM) collaborating with the Australian Bureau of Statistics (ABS) concerning methods to produce reliable estimates for small areas from sample survey data.*

SCHOOL SEGREGATION, CLASS SIZE AND STUDENT ACHIEVEMENT PATTERNS IN SALVADOR DE BAHIA (BRAZIL)

Paulo A. Meyer M. Nascimento

Brazilian National Institute for Applied Economic Research – IPEA
SQS 402, bloco J, ap. 108, Brasília-DF, 70236-100, BRAZIL
Paulo.nascimento@ipea.gov.br

This paper discusses the endogeneity problem in the estimation of class size effects and estimates a multilevel value-added model for Salvador, the capital city of the Brazilian State of Bahia. The endogeneity problem is a common methodological problem in the estimation of relationships between school resources (such as class size) and student achievement. In the case of class sizes, it arises because students are not randomly assigned into classes – class composition is rather the result of optimization behaviours among parents, school principals, teachers and policy makers. This study details the methodological problems related to omitted variable bias that is likely to occur when resources are endogenously determined – and describes ways to construct instrumental variables (IVs) to identify and estimate the exogenous part of resource variation. However, no appropriate IVs could be applied to the available data set. This was due to the lack of longitudinal data or institutional rules limiting class sizes exogenously. Thus a multilevel value-added model is applied to cross-sectional data for a sample of schools, classes and pupils in an attempt to identify factors associated with achievement in Salvador. Although no causal links could be examined, the results suggest some sorting patterns in terms of achievement and class sizes – with public school students studying in larger classes and scoring less than their peers in the private sector.

References:
Angrist, J. D. and Krueger, A. B. (2001), 'Instrumental variable and the search for identification: from supply and demand to natural experiments'. Journal of Economic Perspectives, 15, 69-85.
Averett, S. L. and McLennan, M. C. (2004), 'Exploring the effect of class size on student achievement: what have we learned over the past two decades?' In G. Johnes and J. Johnes (eds), International Handbook on the Economics of Education. Cheltenham: Elgar.
Goldstein, H. (1995), Multilevel Statistical Models. (2nd ed.). London: Edward Arnold.

Paulo A. Meyer M. Nascimento *currently works as a research officer at the Brazilian National Institute for Economic Applied Research – IPEA, in Brasília (capital of Brazil). His areas of interest are Education assessment, Human Capital formation and Labour markets. Paulo has been working with education policy evaluation, specialised labour shortages in Brazil (e.g. engineers) and R&D investments by Brazilian companies. Paulo is also regularly involved in training on Economics, as well as on causal inference and spurious relationships, for journalists and other professionals interested in Economics and Statistics.*

PERCEPTION OF HEALTH RISKS BY THE USE OF PESTICIDES IN THE RING OF "CENOTES" (SINKHOLES) IN THE STATE OF YUCATAN, MÉXICO

Jorge Navarro-Alberto[1], *Angel Polanco-Rodríguez*[2]

[1] *Departamento de Ecología. Campus de Ciencias Biológicas y Agropecuarias, UADY. Km. 15.5 Carr. Merida-Xmatkuil. CP. 97000. Merida, Yucatan, Mexico*
jorge.navarro@uady.mx

[2] *Depto. de Medicina Social y Salud Pública. Centro de Investigaciones Regionales, UADY. Av. Itzaes No. 490 x 59-A. CP 97000. Merida, Yucatan, Mexico*
polanco07@gmail.com

The pressure of extensive farming activities in scenarios surrounded by poverty and underdevelopment, seeking to improve socio-economic conditions, has impacted the aquifer in the Yucatan Peninsula, Mexico. The aquifer in that region is catalogued as "vulnerable" as a consequence of its karstic origin, in which fractures, dissolution channels and presence of sinkholes allow fast infiltration of pollutants spread on the soil. Recently, an increase in the number of cases of cancers for people living near sinkholes in Northern Yucatan has been found, particularly in an area known as the Ring of "Cenotes". This contribution describes the first results of an ongoing research project seeking to determine the factors associated to the increasing number of cancers in the Ring of "Cenotes", mainly those linked to the use and management of groundwater and pesticides. Following the suggestions given by White et al. (2005), surveys were applied (by a two-stage cluster random sampling scheme) to people living in eleven municipalities in the study area, in order to obtain information related to their socio-cultural background, management of agrochemicals associated to farming activities, how people perceive health risks by potential or actual pollutants, and the way water resources are used. Results of these surveys will be presented and potential biases of the information gathered from the questionnaires will also be discussed.

References:
White, P.C.L., Vaughan Jennings, N., Renwick, A.R., and Barker, N.H.L. (2005). Questionnaires in ecology: a review of past use and recommendations for best practice. Journal of Applied Ecology. 42: 421-430.

Jorge Navarro-Alberto *(PhD Otago) is currently the Head of the Departamento de Ecología and a Senior Lecturer at the Universidad Autonoma de Yucatan, in Merida, Yucatan, Mexico. His area of interest is the design and statistical analysis of ecological and environmental studies in the Tropics, particularly in estuaries, coastal lagoons and sinkholes in Southeastern Mexico. Jorge has been involved extensively in statistical consultation, and he has also been working in research on novel statistical methodologies for the analysis of species co-occurrences in community ecology.*

A STATISTICAL METHOD FOR ANALYSIS OF GENE EXPRESSION DATA FROM RT-QPCR

Teresa Neeman[1], *Victoria Ludowici*[2]

[1] *Statistical Consulting Unit, John Dedman Building*
Australian National University, Canberra ACT 0200
teresa.neeman@anu.edu.au

[2] *Plant Science Division, Research School of Biology*
46 Biology Place, Australian National University, Canberra ACT 0200
victoria.ludowici@anu.edu.au

Gene expression analysis is de rigueur in molecular biology research, with real-time quantitative reverse transcription polymerase chain reaction (RT-qPCR) the most widely used technique for quantifying and comparing selected gene expression across genotypes or between treatments. The semi-quantitative assays of yesterday have been usurped by methods that promise quantitative results that can be compared statistically. However, statistical methods for designing and analysing RT-qPCR experiments are still evolving. Variability across runs and across samples due to sample preparation and assay materials require that good experimental design be carefully considered. Relative gene expression estimates are sensitive to estimates of amplification efficiency which must be estimated either from the data or from a separate dilution experiment. It is well known that small errors in amplification efficiency estimation can result in large errors in relative expression estimates and grossly affect inference. In this talk we review the standard assay protocol, the statistical design and analysis issues in a typical relative gene expression experiment. The data example will look at an experiment of biological samples tested under 8 treatments, a dozen genes of interest and a reference gene for normalisation.

Terry Neeman *currently works as a statistical consultant with the ANU Statistical Consulting Unit at the ANU in Canberra. Her area of interest is the application of good statistical methodology in the biological sciences. Terry has been involved extensively in consultation and the development of training programs in the design and analysis of lab-based experiments.*

TESTING THE VALIDITY AND RELIABILITY OF STUDENT EVALUATION OF TEACHING IN ACCOUNTING AND FINANCE PROGRAMS

Foon Lee Ng[1] and Welman Tan Lay Khong[2]

School of Accounting & Finance, Taylor's University College, Malaysia
Taylor's University College, Taylor's Lakeside Campus
No.1, Jalan Taylor's, 47500 Subang Jaya Selangor Darul Ehsan, Malaysia
[1]foonlee.ng@taylors.edu.my
[2]welman.tan@taylors.edu.my

Most universities are using the Student Evaluation of Teaching (SET) as an instrument for students to assess a lecturer's teaching performance. It is an essential instrument to reflect the feedback in enhancing the quality of teaching and learning. The purpose of this paper is to test and assess the reliability and validity of the SET as an instrument in evaluating teaching and learning outcomes of the accounting and finance program at the university.

A study was conducted among 200 students in the accounting and finance twinning programs with universities from Australia and UK at Taylor's University College. Students were required to evaluate the teaching and learning experiences during their course of study by answering 10-item questionnaires:

1. The outline and expectations for this course as supplied by the teacher were clear;
2. The lessons were organised and prepared;
3. The teacher was knowledgeable about the course content;
4. The course content was effectively presented;
5. Opportunities were provided for student participation;
6. The homework and classroom assignments were helpful;
7. The textbooks and/or recommended materials were useful;
8. The teacher was available for consultation and was helpful;
9. The assessment was fair;
10. This course met my needs and goals for future study and/or employment; on a 5-point Likert scale, ranging from 1 to 5, where 1 indicated 'Strongly Disagreement' and 5 indicated 'Strongly Agreement'.

The Cronbach's alpha coefficient was used to test the reliability of the SET by measuring internal consistency. The Cronbach's alpha of 0.906 for the 10-item scale indicated high reliability and internal consistency. Exploratory Factor Analysis (EFA) was performed by applying a Principal Component Analysis. Two factors were extracted, which were subsequently named as 'lecturer attributes' and 'module attributes'. However, the EFA suggested that Question 7 was redundant and it was removed. Separate Confirmatory Factor Analysis (CFA) using AMOS software also confirmed the two factor model. The two factor model was validated using 1000 Bootstrap samples in AMOS.

References

Kember, D., and Doris Y.P. Leung. (2008). Establishing the validity and reliability of course evaluation questionnaires. Assessment & Evaluation in Higher Education, 33 (4), pp. 341-353.
Shevlin, M., Banyard, P., Davies, M. & Griffiths, M. (2000). The Validity of Student Evaluation of Teaching in Higher Education: love me, love my lectures? Assessment & Evaluation in Higher Education, 25 (4), pp. 397-405.
Wilson, Keithia L., Lizzio, Alf and Ramsden. (1997). The Development, Validation and Application of the Course Experience Questionnaire. Studies in Higher Education, 22 (1), pp. 33-53

Foon Lee Ng *is a Statistics lecturer at Taylor's University College, Malaysia. She teaches mostly statistics subjects/modules in Accounting & Finance twinning programs with universities from Australia and the UK. Her research interests are in applied statistic, quantitative methodology and issues of teaching and learning.*

FAST PRIOR SENSITIVITY ANALYSIS IN BAYESIAN DNA SEQUENCE SEGMENTATION MODELLING

Darfiana Nur[1], Kerrie L. Mengersen[2] and Judith Rousseau[3]

[1]School of Mathematical and Physical Sciences, The University of Newcastle
University Drive, Callaghan, NSW 2308, Australia
Darfiana.Nur@newcastle.edu.au

[2]School of Mathematics and Statistics, Queensland University of Technology
Gardens Point, Brisbane, QLD4001, Australia
k.mengersen@qut.edu.au

[3]Universite Paris Dauphine
Place du Marchal De Lattre de Tassigny, Paris, France, 75016
rousseau@ceremade.dauphine.fr

We consider the problem of conducting comprehensive sensitivity analysis for complex Bayesian models and examine an importance sampling approach based on a single MCMC chain. We use the method to assess sensitivity of priors in a Bayesian hidden Markov model for homogeneous segments of DNA sequences, specifically the genome of the bacteriophage lambda, a parasite of the intestinal bacterium Escherichia coli. The importance sampling approach facilitates fast, comprehensive sensitivity analysis, leading to a better understanding of complex models.

References

Boys, R.J, Henderson, D.A and Wilkinson, D.J. (2000). Detecting homogeneous sements in DNA sequences by using Hidden Markov models. Applied Statistics Vol 49, 269-285.
Boys, R.J and Henderson, D.A. (2004). A Bayesian approach to DNA sequence segmentation. Biometrics Vol 60, 573-588.

Darfiana Nur *received her PhD in Statistics in 1999 at the Curtin University of Technology, Australia. She is Lecturer in Statistics at the School of Mathematical and Physical Sciences, University of Newcastle, Australia. Her research interests include nonlinear time series analysis and modelling, Markov chain Monte Carlo (MCMC) convergence and DNA sequence modelling. Her current research interest in DNA sequence modelling is focusing on research to develop theory, methods and applications of semiparametric Hidden Markov Models (HMMs) in a Frequentist and/or Bayesian frameworks.*

METHOD OF MULTIPLE IMPUTATION FOR CANONICAL CORRELATION ANALYSIS WITH MISSING OBSERVATIONS

Toru Ogura[1], Shin-ichi Tsukada[2], Toshinari Kamakura[3]

[1] Chuo University
1-13-27, Kasuga, Bunkyo-ku, Tokyo, 112-8551, JAPAN
ogura@indsys.chuo-u.ac.jp

[2] Meisei University
2-1-1, Hodokubo, Hino-shi, Tokyo, 191-8506, JAPAN
tsukada@ge.meisei-u.ac.jp

[3] Chuo University
1-13-27, Kasuga, Bunkyo-ku, Tokyo, 112-8551, JAPAN
kamakura@indsys.chuo-u.ac.jp

The canonical correlation analysis is a linear combination of the original variables in each data set to maximize the correlation between the two linear combinations. These linear combinations become the first coordinates in two new systems of coordinates. A second linear combination is then obtained in each data set, subject to the condition that it is uncorrelated with the first linear combination. The procedure is continued until the two new coordinate systems are completely specified.

We cannot use the canonical correlation analysis when there is missing observations. But, in real data, missing observations often appear. Therefore, we study method of using canonical correlation analysis in that case. One method is to exclude the missing observations, but the observation value is not effectively used. Therefore, we propose the method of multiple imputation for canonical correlation analysis with missing observations. And, we propose some multiple imputation methods and compare them by using real data.

References

Kenward, M, G. (2007). Multiple imputation: current perspectives. Statistical Methods in Medical Research, 16, 199-218.

Peng, Y., Zhang, D., and Zhang, J. (2010). A New Canonical Correlation Analysis Algorithm with Local Discrimination. Bioinformatics, 31, 1-15.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, RB. (2001). Missing value estimation methods for DNA microarrays. Bioinformatics, 17, 520-525.

Toru Ogura *currently works as a Assistant Professor with the Chuo University in Japan. My area of interest is the canonical correlation analysis.*

STATISTICAL CLASSIFICATION ALGORITHM FOR INFRASONIC FREQUENCY MULTIVARIATE TIME-SERIES DATA WITH APPLICATION TO HUMAN ACTIVITY RECOGNITION

Kosuke Okusa[1], *Toshinari Kamakura*[2]

[1] *Graduate School of Science and Engineering, Chuo University*
1-13-27 Kasuga, Bunkyo-ku, Tokyo
k.okusa@me.com

[2] *Department of Science and Eigineering, Chuo University*
1-13-27 Kasuga, Bunkyo-ku, Tokyo
kamakura@indsys.chuo-u.ac.jp

Human activity recognition is an important task which has many potential applications (Gafurov et al. (2006), Yun et al. (2006) etc.). In recent years, researchers have become interested in deploying on-body acceleration/angular speed sensors to collect observations and applying machine learning techniques (e.g. Hidden Markov Model (HMM), Support Vector Machine (SVM)) to model and recognize activities (Kunze and Lukowicz (2008)). However, such methods only detect simple activity (to detect pace of walking, standing, sitting). We propose a new statistical detection algorithm for complex human activities based on the observations obtained by acceleration/angular speed sensors.

Two obstacles have to be overcome for detecting human activities. Firstly, although observations of human motion are mainly based on infrasonic frequency (0.1-2Hz), real-time feedback analysis requires short time window size and only a fraction of the wave can be utilized.

Secondly, we must adjust individuality parameters for efficient recognition of human activities among many individuals. Even if examinees do the same motion at the same pace, it includes many differences of such parameters.

To solve these problems, we propose an infrasonic frequency-phase analysis method (we call "Short-Time Wavelet Transform") and multivariate time-series/high-dimensional feature quantities classification algorithm. As for examples, we measure 17-categorized weight training equipment and 32-categorized aerobics exercises, and verify recognition rate of our algorithm. We illustrate the better performance of our algorithm in comparison to other methods.

References

Gafurov, Davrondzhon., Helkala, Kirsi. and Søndrol, Torkjel
(2006). Biometric Gait Authentication Using Accelerometer Sensor. Journal of Computers, 1-7, 51-59.
Yun, Jaeseok., Patel, Shwetak., Reynolds, Matt and Abowd, Gregory.
(2008) Quantitative Investigation of Inertial Power Harvesting for Human-powered Devices. UBICOMP 2008, 74-83.
Kunze, Kai., Lukowicz, Paul.
(2008). Dealing With Sensor Displacement In Motion-Based On-body Activity Recognition Systems. UBICOMP 2008, 20-29.

Kosuke Okusa *currently works as a Research Assistant and Graduate Student with the Statistical Data Analysis Laboratory, Graduate School of Science and Engineering, Chuo University, Kourakuen Campus, in Tokyo. His area of interest is the Pattern Recognition of Human Avtivity, Human Motion Analysis and Authentication, Statistical Motion Picture Analysis.*

SO HOW MANY SPECIES ARE THERE REALLY? EXPERT ELICITATION OF BIODIVERSITY ON CORAL REEFS

Rebecca A. O'Leary[1], Rebecca Fisher[2], Samantha Low Choy[3], Kerrie Mengersen[4], Julian Caley[5]

[1] Australian Institute of Marine Science, The UWA Oceans Institute (M096)
35 Stirling Highway, Crawley WA 6009
r.oleary@aims.gov.au

[2] Australian Institute of Marine Science, The UWA Oceans Institute (M096)
35 Stirling Highway, Crawley WA 6009
r.fisher@aims.gov.au

[3] School of Mathematical Sciences, Queensland University of Technology,
GPO Box 2434 Brisbane QLD 4001, Australia
s.lowchoy@qut.edu.au

[4] School of Mathematical Sciences, Queensland University of Technology
GPO Box 2434 Brisbane QLD 4001, Australia
k.mengersen@qut.edu.au

[5] Australian Institute of Marine Science
PMB No. 3, Townsville, QLD 4810, Australia
j.caley@aims.gov.au

In the ecological field, expert opinion has been acknowledged as providing valuable information in modelling, particularly when the observed data are limited or unreliable. A convenient framework for combining expert information with observed data is Bayesian statistical modelling (O'Hagan et al. 2006; Low Choy et al. 2009).

The aim of this research is to estimate the total number of species on coral reefs globally. Surprisingly, there are very limited data on the number of species on coral reefs. Although the number is thought to lie somewhere between 1 and 9 million, the World Register of Marine Species database (the most authoritative list of marine species globally) currently contains around 200,000 species, and only 3% of these records can currently be identified as belonging to coral reefs. Here we design an expert elicitation approach to ask taxonomists to estimate the number of species and uncertainty within their taxonomic group.

The total number of species is made up of three unknown components: named species, discovered but un-named species and completely undiscovered species. For each component we elicited from the expert: the lower and upper bound, sureness (how sure the real number lies in these bounds) and the best guess of the most likely value of the number of species. This information was then used to estimate the parameters of either a normal or log-normal distribution. From this the "mathematical" best guess (mode) can be calculated, which is subsequently shown to the expert as feedback in the form of a boxplot, displaying their initial estimates plus the mathematical best guess and confidence bound with a different certainty to the original sureness. The interactive elicitation questionnaire is written in tcltk in R. The results demonstrate how expert elicitation can be used to estimate parameters in ecology that would otherwise remain elusive to traditional data driven techniques.

References

Low Choy, S., O'Leary, R. A., and Mengersen, K. (2009). Elicitation by design in ecology: using expert opinion to inform priors for Bayesian statistical models. Ecology, 90, 265-277.
O'Hagan, A., C. E. Buck, A. Daneshkhah, J. R. Eiser, P. H. Garthwaite, D. J. Jenkinson, J. E. Oakley, and T. Rakow (2006). Uncertain Judgements: Eliciting Expert Probabilities. United Kingdom: Wiley.

Rebecca O'Leary *currently works as Biostatician for the Australian Institute of Marine Science (AIMS). Rebecca completed her PhD (2008) at Queensland University of Technology, on the use and impact of expert knowledge on models for rare events. Her primary area of research is in medical, ecological and environmental modelling, in both frequentist and Bayesian contexts. Specific methodological interests are in Bayesian statistics, expert elicitation and hierarchical modelling. She is Vice-President of the Western Australia branch of the Statistical Society of Australia.*
Rebecca Fisher *completed her PhD in larval ecology (2002) at James Cook University. Currently a postdoctoral fellow with CReefs node of Census of Marine Life at AIMS, she is developing models for estimating global species richness on coral reefs and ascertaining the proportion of these species that are dependent on coral reef habitat. She has also been developing a novel semi-automated approach for modelling spatial and taxonomic patterns in research effort on coral reefs.*

CONVERGENCE RATE OF WAVELET EXPANSIONS OF STOCHASTIC PROCESSES

Yuriy Kozachenko[1], Andriy Olenko[2], Olga Polosmak[3]

[1] Department of Probability Theory, Statistics and Actuarial Mathematics
Kyiv University, Kyiv 01601, Ukraine
ykoz@ukr.net

[2] Department of Mathematics and Statistics, La Trobe University, Victoria, 3086, Australia,
a.olenko@latrobe.edu.au

[3] Department of Probability Theory, Statistics and Actuarial Mathematics
Kyiv University, Kyiv 01601, Ukraine
DidenkoOlga@yandex.ru

In various statistical, data compression, signal processing applications and simulation the problem of analyzing a continuous-time random process could be converted to that of analyzing a random sequence, which is much simpler. Multiresolution analysis provides an efficient framework for the decomposition of random processes. This approach is widely used in statistics to estimate a curve given observations of the curve plus some noise. Various extensions of the standard statistical methodology were proposed recently. These include curve estimation in the presence of correlated noise. For these purposes the wavelet based expansions have numerous advantages over Fourier series, and often lead to stable computations.

However, in many cases numerical simulation results need to be confirmed by theoretical analysis. Recently, considerable attention was given to the properties of the wavelet transform and the wavelet orthonormal series representation of random processes.

We focus our attention on uniform convergence of wavelet expansions for random processes.
We consider random processes $\mathbf{X}$ (t) and their approximations by sums of wavelet functions, as follows:

$$X_{n,k_n}(t) := \sum_{|k|\le k_0} \xi_{0k}\varphi_{0k}(t) + \sum_{j=0}^{n-1}\sum_{|k|\le k_j} \eta_{jk}\psi_{jk}(t),$$

where $k_n := (k_0,...,k_{n-1})$.

We show that, under suitable conditions, the sequence $X_{n,k_n}(t)$ converges in probability in Banach space $C([0,T])$, i.e.

$$P\{\sup_{t\in[0,T]} \left|X(t) - X_{n,k_n}(t)\right| > \varepsilon\} \to 0,$$

when $n \to \infty$ and $k_n \to \infty$ for all $j \in \{0,1,...\}$.

The numbers *n* and k_n of terms in the truncated series $X_{n,k_n}(t)$ can approach infinity in any arbitrary way.

One sees that we have the most general class of such wavelet expansions in comparison with particular cases considered by other authors.

The rate of convergence is another natural question which would be very useful to answer in various computational applications, especially if we are interested in the optimality of the stochastic approximation or the simulations. This question has not been addressed yet. We have obtained the exponential rapidity of convergence:

$$P\{\sup_{t\in[0,T]} \left|X(t) - X_{n,k_n}(t)\right| > u\} \le 2\exp\{-\frac{(u-\sqrt{8u\delta(\varepsilon_{k_n})})^2}{2\varepsilon_{k_n}^2}\},$$

where $u > 8\delta(\varepsilon_{k_n})$. Explicit expressions for constants were obtained.

This work was partly supported by La Trobe University Research Grant *"Stochastic Approximation in Finance and Signal Processing"*.

Andriy Olenko *currently works as a Lecturer at La Trobe University, Melbourne. His areas of interest are Spatial Statistics, Random Fields, Limit Theorems, Stochastic Approximation.*

SKEW-NORMAL VARIATIONAL APPROXIMATIONS

John T. Ormerod

School of Mathematics and Statistics, University of Sydney
Sydney 2006, Australia
john.ormerod@sydney.edu.au

Variational approximations facilitate approximate Bayesian inference for complex statistical models and are emerging as fast, deterministic alternatives to Monte Carlo methods. In this talk we introduce skew-normal variational approximations. This type of variational approximation draws upon the work of Ormerod & Wand (2010) by using the multivariate skew-normal density of Azzalini & Dalla Valle (1996) to approximate posterior densities. The method will be placed into context by highlighting similarities with other deterministic methods including the Laplace's method, variational Bayes and Gaussian variational approximations. We will use one or two biomedical applications, time permitting, to illustrate the superior accuracy of skew-normal variational approximations over these other methods.

References

Azzalini, A. and Dalla Valle, A. (1996). The multivariate skew-normal distribution. Biometrika, 83, 715-726.

Ormerod, J.T., and Wand, M.P. (2010). Gaussian variational approximate inference for generalized linear mixed models. Submitted to Journal of Computational and Graphical Statistics.

John Ormerod *is a recently appointed a Lecturer in Statistics at the University of Sydney. His areas of interest include Variational Approximations, Generalised Linear Mixed Models and Semiparametric Regression. John has recently published several papers with his PhD supervisor Matt Wand developing variational approximations for statistical audiences.*

USE OF STOCHASTIC DIFFERENTIAL EQUATIONS FOR ESTIMATION IN HYDROLOGY

Daniel E. Pagendam

CSIRO Mathematics, Informatics and Statistics
120 Meiers Rd, Long Pocket, Queensland 4068
Dan.Pagendam@csiro.au

Stochastic differential equations (SDEs) are routinely employed in fields such as economics and finance as statistical models, however, their use in surface water hydrology is uncommon. There appears to be even less use of these methods for the statistical analysis of hydrological data. One of the potential drawbacks of using SDEs in hydrology, as opposed to finance, is the difficulty in formulating a suitable parametric model for the process under investigation. This involves defining the drift and volatility behaviour of the process.

In recent years, there has been a growing interest in the use of nonparametric statistical methods for estimating the drift and volatility functions of stochastic differential equations. Recently Bandi and Philips (2003) and Bandi and Moloche (unpublished) showed that in order to estimate these functions nonparametrically, one need only assume that the underlying stochastic process is Harris recurrent. Harris recurrence simply requires that the continuous trajectory of the process visits those sets in the state space, having non-zero probability measure, infinitely often. This is much weaker than assuming stationarity or mixing and can in fact accommodate non-stationary processes (commonplace in hydrology where data shows seasonality).

Identifying the drift and volatility functions nonparametrically relies on data sets where measurements are made with relatively high frequency (e.g. daily) so that changes in the process can be identified over small intervals of time. The drift and volatility functions are identified using Nadaraya-Watson estimators and we show how approximate diffusion bridges (Bladt and Sørensen, 2007) between the daily observations can be used to put confidence bounds on quantities such as total discharge from a stream. Furthermore, the method is particularly useful for dealing with ‘gappy’ time-series and time-series subject to left-censoring.

References

Bandi, F.M. and Moloche, G. (unpublished manuscript). On the functional estimation of multivariate diffusion processes.

Bandi, F.M. and Phillips, P.C.B. (2003). Fully nonparametric estimation of scalar diffusion models. Econometrica, 71, 241-283.

Bladt, M. and Sørensen, M. (2007). Simple simulation of diffusion bridges with applications to likelihood inference for diffusions. Centre for Analytical Finance, University of Aarhus. Working Paper Series, No. 225.

Daniel Pagendam *is an OCE postdoctoral research fellow with CSIRO Mathematics, Informatics and Statistics. He completed a Bachelor's degree in environmental science (Hons. I) in 2002 before commencing an M.Sc. and PhD in statistics at the University of Queensland. Daniel has previously worked in water quality modelling with the Queensland government and the CRC for Catchment Hydrology. His current employment with CSIRO is focussed on developing novel statistical methodology for analyzing water quality and hydrological data.*

GENERALIZED POISSON LAWS ARE POISSON MIXTURES

Anthony G Pakes

University of Western Australia
School of Mathematics and Statistics, UWA
35 Stirling Highway, Crawley, WA 6009
pakes@maths.uwa.edu.au

Generalized Poisson laws fit many data sets; see Consul (1989). Joe and Zhu (2005) present a not entirely compelling proof that generalized Poisson laws are Poisson mixtures. By showing that the Lambert W function is the Laplace exponent of a positive infinitely divisible law I can give a compelling proof of the Poisson mixture property. This approach shows also that generalized Poisson laws are generalized negative-binomial convolutions.

References

Consul, P.C. (1989). Generalized Poisson Distributions: Properties and Applications. New York: Marcel Dekker, Inc.

Joe, H., and Zhu, R. (2005). Generalized Poisson distribution: the property of mixture of Poisson and comparison with negative binomial distribution. Biometrical J. 2, 219-229.

Tony Pakes *currently holds an academic appointment at the University of WA. This year his teaching covers time series analysis, Bayesian inference, Markov chains and processes, and the principles and applications of stochastic calculus. Current research interests include branching processes, extreme value theory, structure of statistical distributions etc.*

RAINFALL DEPTH AREA CURVES DERIVED FROM A SPATIAL-TEMPORAL MODEL FOR EXTREME RAINFALL

Palmer, M.J

CSIRO, Mathematics, Informatics and Statistics
Leeuwin Centre, 65 Brockway Rd, Floreat WA, 6014
Mark.palmer@csiro.au

Spatial–temporal models developed for analysis of extreme rainfall have proven effective at characterizing the characteristics of rainfall at a site, both for gauged and ungauged locations. However, they have not been as useful for characterizing the extremes of rainfall over an area, such as a catchment. This is apparently due to incomplete modelling of dependence (Sang and Gelfand, 2009). An approach, based on copulas, has been developed that attempts to overcome this problem and its application to a study area in NSW will be described.

References

Sang, H., and Gelfand, A. E. (2009), Continuous spatial process models for spatial extreme values. Journal of Agricultural, Biological, and Environmental Statistics, 15, 49-65.

Mark Palmer *is in the Environmental Informatics program of the CSIRO Division of Mathematics, Informatics and Statistics, and located in Perth. His main interests are in spatial statistics within a Bayesian framework, with applications to problems associated with water, including monitoring quality, varied aspects of rainfall and sediment composition.*

MUTUAL INFORMATION AND KERNEL METHODS FOR THE INFERENCE OF GENETIC REGULATORY NETWORKS

Chris Pardy[1], *Susan Wilson*[2], *Sally Galbraith*[3]

[1] *Prince of Wales Clinical School, University of New South Wales*
Level 4 Lowy Cancer Research Centre, Sydney NSW 2052
cpardy@unsw.edu.au

[2] *School of Mathematics and Statistics & Prince of Wales Clinical School The University of New South Wales, Sydney NSW 2052; Centre for Mathematics and its Applications, ANU*
sue.wilson@anu.edu.au

[3] *School of Mathematics and Statistics, University of New South Wales*
Sydney NSW 2052
sally.galbraith@unsw.edu.au

Experimental work in the field of genetics creates large quantities of high-dimensional data. The systems biology approach attempts to integrate the multiple sources of these data: including gene expression levels, single nucleotide polymorphisms (SNPs) and clinically measured phenotypic outcomes. A natural way to express the relationship between these data is via a network of associations. Information theoretic and machine learning techniques such as mutual information (MI) and clustering can be used to identify properties of these networks, describe biological systems and determine potential targets for novel medical therapies.

We extend a previous approach (Zhang and Horvath, 2005) by using MI as a measure of association that is valid for both continuous and discrete variables. This allows us to include data with any distribution in a single network. The network can be grouped according to association with important clinical measurements with the aim of identifying modules containing biological pathways and highly connected "hub" genes. Various clustering techniques have been explored for the identification of such modules. Kernel density estimation has been used to develop non-parametric estimators for the MI between gene expression levels and SNPs. The gene expression levels are modelled as a mixture distribution where the continuous gene expression levels are mixed with probabilities given by the discrete SNP distribution, leading to a joint distribution with both continuous and discrete parts. We make no parametric assumptions regarding the distribution of the continuous variables.

References

Zhang, B. and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. Statistical Applications in Genetics and Molecular Biology, 4(1), 1128.

Chris Pardy *is a PhD student at the UNSW faculty of medicine with a background in mathematical statistics and professional experience in clinical trials. He is currently interested in the application of information theory, machine learning and rigorous probability to problems in bioinformatics.*

GENERALIZED LEAST SQUARES METHOD IN GSTAR(1,1) WITH SPATIALLY CORRELATED ERRORS

Udjianna Pasaribu[1], Nunung Nurhayati[2], Khreshna Syuhada[3]

[1] Statistics Research Division-Institut Teknologi Bandung (ITB), Jalan Ganesa 10 Bandung, udjianna@math.itb.ac.id

[2] Department of Mathematics and Natural Sciences-Universitas Jenderal Soedirman, Jalan dr. Soeparno, Karangwangkal Purwokerto nunung90@yahoo.com

[3] Statistics Research Division-Institut Teknologi Bandung (ITB), Jalan Ganesa 10 Bandung, khreshna@math.itb.ac.id

Paramater estimation in a Generalized Space Time Autoregressive (1,1) or GSTAR (1,1) model through a Generalized Least Squares method is applied by assuming that the errors between locations are dependent. We consider that the error is influenced linearly by nearby errors. By assumming that the process is stationary and the errors follow a difference Martingale process, it is shown analytically that the estimator vector is unbiased and consistent.

References

Anselin, L. (2003). Spatial Econometrics. Blackwell Publishing Ltd

Nurani, B. (2002). Suatu Model Generalisasi Space-Time Autoregresi dan Penerapannya pada Produkai Minyak Bumi. Dissertation of Doctoral Program. Bandung: Institut Teknologi Bandung.

Nurhayati, N., Pasaribu U.S., Darwis, S. (2006). Simulation Studies of the Least Square Estimator Convergence in Generalized Space Time Autoregressive Model. Proceeding of National Conference on Mathematics XIII, UNESA, Semarang. 509-515.

Udjianna S. Pasaribu *completed a Ph.D Degree from University of Wales, Swansea, United Kingdom in 1993. Udjianna has worked as a Lecturer at Institut Teknologi Bandung (ITB), Indonesia, since 1987. Her research areas include Actuarial Sciences, Space-Time Modelling, and Mortality Tables.*

A MODEL-BASED APPROACH TO DESIGNING A FISHERY INDEPENDENT SURVEY

David Peel[1], Mark Bravington[1], Natalie Kelly[1], Simon Wood[2], Ian Knuckey[3]

[1] Wealth from Oceans National Research Flagship and CSIRO Mathematics, Informatics & Statistics
Castray Esplanade, Hobart TAS 7001 Australia
David.Peel@csiro.au

[2] Mathematical Sciences, University of Bath, Bath, BA2 7AY United Kingdom
s.wood@bath.ac.uk

[3] Fishwell Consulting, 22 Bridge St, Queenscliff VIC 3225, Australia
ian@fishwell.com.au

This talk describes a model-based approach to survey design for wildlife abundance estimation. Managing wild animal populations requires high quality data on their abundance and distribution, and a great deal of consideration can go into designing and running surveys for that purpose. Marine fish stocks can be particularly difficult to monitor, as they cannot be counted directly and their spatial distributions can vary greatly over time. Fishery independent surveys (FIS) are one of the most valuable tools for fish stock assessment, and underpin the provision of management advice in numerous major ground fish fisheries. Specifically, this talk describes the use of Generalized Additive Models to evaluate model-based designs for wildlife abundance surveys where substantial pre-existing data are available. These data are often available in fisheries where historical data of fishermen catches exist. Compared to conventional stratified designs or design-based designs, our model-based designs can be both efficient and flexible, for example in allowing uneven sampling due to survey logistics, and provide a general framework to answer specific design questions. As an example, we shall briefly present the design and implementation of a trawl survey for eleven fish species along the continental slope off South-East Australia.

David Peel *currently works for CSIRO as part of the CSIRO Mathematical and Information Science (CMIS) group. In this role he generally works on quantifying fish and marine mammal movement, spatial distribution and abundance. His current areas of interests are model based survey design, spatial modelling, hidden Markov models, line transect surveys and developing camera based aerial survey tools.*

SEXUAL ORIENTATION DATA IN OFFICIAL PROBABILITY SURVEYS: FINDINGS FROM THE SEXUAL ORIENTATION DATA COLLECTION STUDY, NEW ZEALAND

Frank Pega[1], Alistair Gray[2], Jaimie F. Veale[3]

[1] Independent Researcher
127 Eden Street, Island Bay, Wellington 6023, New Zealand
fpega@hotmail.com

[2] Director, Statistics Research Associates Ltd
8 Bristol Street, Island Bay, Wellington 6023, New Zealand
alistair@statsresearch.co.nz

[3] Researcher, School of Psychology, Massey University
Private Bag 102- 904, North Shore Mail Centre, Auckland 0745, New Zealand
j.f.veale@massey.ac.nz

Common standards for official statistics are being adopted on both sides of the Tasman, particularly where there are emerging categories, e.g. the ANZSCO for occupation (Australian Bureau of Statistics and Statistics New Zealand, 2009). Sexual orientation emerges as an important official social statistic in New Zealand (Statistics New Zealand, 2008). We conducted the Sexual Orientation Data Collection Study on behalf of the Ministry of Social Development and its partners Statistics New Zealand and the Ministry of Health in New Zealand. The study aimed to develop standards for sexual orientation data collection in official New Zealand surveys. We reviewed literature, conducted focus groups and key informant interviews with sexual minority people as well as producers and consumers of official statistics, and analysed existing official sexual orientation data. A 15-member panel of national and international experts provided advice and peer-review throughout the research project. The Conceptual Framework of Sexual Orientation (Pega, Gray and Veale, 2010) proposes working definitions of sexual orientation concepts; describes dimensions of the concepts; discusses variables framing the concepts; and outlines conceptual grey areas. The Sexual Orientation Measurement and Data Collection Framework (Pega, Gray and Veale, 2010) identifies and discusses key methodological issues in relation to sexual orientation; synthesises existing evidence on each issue; reviews international best practice in relation to each issue; proposes strategies to address each methodological issue in the New Zealand context, including preliminary standard questions for inclusion in surveys, standard modes for survey administration, and standards for data disaggregation and analysis. If producers of official statistics apply these standards in their surveys, high-quality official data on sexual orientation will become available for developing, costing, and evaluating public policies and services targeted at populations defined by sexual orientation.

References

Australian Bureau of Statistics and Statistics New Zealand. 2006. ANZSCO - Australian and New Zealand Standard Classification of Occupations, First Edition, Revision 1. ABS cat. no. 1220.0 Australian Bureau of Statistics, Canberra, Australia.

Pega, F., Gray, A., and Veale, J. F. 2010, 'Sexual orientation data in probability surveys: Improving data quality and estimating core population measures from existing New Zealand survey data', Official Statistics Research Series, vol. 6, no. 2. http://www.statisphere.govt.nz/official-statistics-research/series/2010/page2.aspx

Statistics New Zealand. 2008, Considering sexual orientation as a potential official social statistic, Statistics New Zealand, Wellington, New Zealand.

Frank Pega *is a public and international health researcher and epidemiologist with an interest in the social determinants of health and health equity. He is currently employed at the Otago University Wellington School of Medicine, after working as a consultant for the World Health Organization (Geneva Headquarters) and the Ministry of Social Development (New Zealand). He presents findings from the Sexual Orientation Data Collection Study, an Official Statistics Research (Statistics New Zealand) project, of which he is the lead researcher.*

ON GENERALIZED FRACTIONAL PROCESSES WITH CONDITIONAL HETEROSCEDASTICITY

Shelton Peiris[1], *G.S.Dissanayake*[2]

[1]*School of Mathematics and Statistics, The University of Sydney*
shelton.peiris@sydney.edu.au

[2]*School of Mathematics and Statistics, The University of Sydney*
dissan@maths.usyd.edu.au

The class of stationary long-memory processes became very popular among researchers due to its applicability in many time series and related problems, especially in finance. This is obtained by allowing the degree of differencing d in $(1 - B)^d$ to take values in (-0.5, 0.5) and fitting an ARMA model (FARIMA) to the differenced data given by $Y_t = (1 - B)^d Xt$. This paper considers an extension of the above fractionally differenced class taking the differenced series to be $Y_t = (1 - \alpha B + B^2)^d X_t$ and fitting an ARMA model with generalised autoregressive conditional heteroscedastic (GARCH) errors. Some sufficient conditions for stationarity, ergodicity and existence of higher order moments are established in terms of Gagenbauer polynomials. Our extension, GAFARIMA(p,d,q) - GARCH(r,s) clearly includes the FARIMA model as a special case. Some simulations results are presented to illustrate the theory.

Shelton Peiris *currently works as Associate Professor in the School of Mathematics and statistics at Sydney University. His research includes statistical analysis in time series analysis and financial econometrics.*

APPLICATIONS OF KUSUOKA REPRESENTATION OF DUAL RISK MEASURES

Darinka Dentcheva[1], Spiridon Penev[2], Andrzej Ruszczyński[3]

[1] Stevens Institute of Technology, Hoboken, NJ,
darinka.dentcheva@stevens.edu

[2] The University of New South Wales, Sydney,
s.penev@unsw.edu.au

[3] Rutgers University, Piscataway, NJ,
rusz@business.rutgers.edu

We develop analytic formulae for a new family of coherent risk measures in L^p spaces, $p > 1$, which are tailored towards applications in Finance. The measures are represented as suprema of integrals of a specific risk measure, the Average Value of Risk, with respect to a suitable convex set of probability measures in (0,1]. Such representations are called Kusuoka representations. Interestingly, when $p=2$ these measures relate to the Fano factor in statistics, which is a specific expression for a noise-to-signal ratio.

We also discuss the problem of inference about risk measures given by their Kusuoka representations. Technically, after replacing expectations by sample averages, calculating the risk becomes a problem of minimizing a convex function subject to convex constraints. Convex programming and stochastic approximation techniques can then be used to find the solution.

References

Dentcheva, D., and Penev, S. (2010). Shape-restricted inference for Lorenz curves using duality theory. Statistics & Probability Letters, 80, 403-412.
Dentcheva, D., Penev, S., and Ruszczński, A. (2010). Kusuoka Representation of Higher Order Dual Risk Measures. To appear, Annals of Operations Research.
Shapiro, A., Dentcheva, D., and Ruszczyński, A. (2009). Lectures on Stochastic Programming: modeling and theory. Philadelphia: SIAM.

Spiridon Penev *currently works at the department of Statistics, The University of New South Wales, Sydney. He has wide-ranging interests in statistical methodology and its applications in diverse areas such as Finance and Risk Management, Behavioral Sciences, and in Engineering.*

SCALED SQUARED DISTANCE-BASED CLASSIFIERS FOR HIGH DIMENSIONAL LOW SAMPLE SIZE DATA

Tung Pham

Department of Mathematics and Applied Statistics, University of Wollongong
NSW, 2522, Australia
tung@uow.edu.au.

A linear classifier is often comprised of a weight vector and a threshold. In high dimensional and low sample size settings, the weight vector often does not reflect the difference between two classes. Some linear classifiers can be transformed into squared distance classifiers. Using Chan and Hall's scale-adjusted method we demonstrate that a classifier based on these scaled squared distances is asymptotically equivalent to the linear classifier with optimal weight vector for high dimensional and low sample size data. We also prove that, under weaker conditions than those in Chan and Hall (2009), the scaled support vector machine classifier enjoys similar properties of the scaled centroid-based classifiers, when the sample sizes diverge at a rate no faster than the tenth root of the dimension.

References

Chan, Y and Hall, Peter (2009). Scaled adjustment for classifiers in high-dimensional low sample size setting. Biometrika 96, 469-478.

Tung Pham *currently works as an associate researcher with Prof Matt Wand at the department of mathematics and applied statistics in the University of Wollongong. His areas of interest are high dimensional data analysis with its applications and generalized linear mixed models.*

USING EDGEWORTH EXPANSION APPROXIMATING TWO- AND THREE-DIMENSIONAL PROBABILITY DISTRIBUTION FUNCTIONS

Margus Pihlak

Tallinn University of Technology, Ehitajate tee 5, 19086 Tallinn
margusp@sta®.ttu.ee

In this talk we present techniques for approximating an unknown distribution function with a well-known and well-studied distribution function. The development of the approximation technique is closely related with the development of matrix algebra. We also present some newer results of matrix algebra. For a more detailed presentation of this kind of matrix algebra see Harville (1997), Pihlak (2004), Kollo and von Rosen (2005), for example. In Pihlak (2008) the two-dimensional distribution function is approximated by means of an Edgeworth expansion. In this presentation we generalize the Edgeworth expansion to the three-dimensional case. This presentation is supported by the Estonian Science Foundation Grant 7656.

References

Harville, A. (1997). Matrix Algebra from a Statistican's Perspective. Springer, New York.
Kollo, T. and von Rosen D. (2005). Advanced Multivariate Statistics with Matrices. Springer, Dordrecht.
Pihlak, M. (2004) Matrix integral. Linear Algebra and Its Applications, , 315-325.
Pihlak, M. (2008) Approximation of Multivariate Distribution Functions. Mathematica Slovaca, 58, 635-652.

Margus Pihlak *currently works as a associated professor at Tallinn University of Technology. His research area concerns application of matrix algebra on multivariate statistical analyzes. He has also worked with dynamical models of ecosystems. Currently he teaches students mathematical statistics and applied probability.*

BANDWIDTH SELECTION FOR KERNEL CONDITIONAL DENSITY ESTIMATION USING THE MCMC METHOD

Julia Polak[1], Xibin Zhang[2], Maxwell L. King[3]

[1] Department of Econometrics and Business Statistics
Monash University, Clayton, Victoria 3800, Australia
julia.polak@buseco.monash.edu.au

[2] Department of Econometrics and Business Statistics
Monash University, Caulfield, Victoria 3145, Australia
xibin.zhang@buseco.monash.edu.au

[3] Office of the PVC (Research & Research Training)
Monash University, Clayton, Victoria 3800, Australia
max.king@adm.monash.edu.au

The availability of an accurate estimator of conditional densities is very important due to the high usage and potential usage of conditional densities in econometrics. Moreover, the conditional density estimator is a generalization of the regression model estimator which is widely used in practice but typically only provides an approximation to the conditional mean. In contrast, the conditional density estimator provides a wider range of properties, therefore allowing the researcher to examine a wider range of hypotheses. For the kernel conditional density estimator we propose an implementation of the MCMC estimation algorithm for optimal bandwidth selection presented by Zhang, King and Hyndman (2006). Our approach is based on the indirect estimation of the conditional density as a ratio of two different joint densities combined with likelihood cross-validation and MCMC sampling to determine the values of bandwidth parameters. In addition, we propose a generalization to the Kullback-Leibler information criterion applicable to conditional density comparison. Our numerical study shows that the MCMC algorithm for bandwidth selection in the kernel conditional density estimator performs much better than the popular normal reference rule for bandwidth selection.

References

Zhang, X., King, M. L. and Hyndman, R. J. (2006). A Bayesian approach to bandwidth selection for multivariate kernel density estimation. Computational Statistics and Data Analysis, 50, 3009-3031.

Julia Polak *graduated with an M.Sc. in Economics from Technion - Israel Institute of Technology in 2007. Her first paper jointly with N. Perach and U. G. Rothblum was published in the International Journal of Game Theory, 2007. This paper is based on her B.A. Project. Currently Julia is conducting Ph.D. research in econometrics at the department of Econometrics and Business Statistics, Monash University, Melbourne. Her research focuses on improvement of kernel conditional density estimation and its application for statistical testing. At the commencement of her Ph.D. research Julia has been awarded the Australian Postgraduate Award, the Dean's Postgraduate Research Excellence Award and the Donald Cochrane Postgraduate Research Scholarship.*

COMPARING SLK AND NAME-BASED STRATEGIES FOR DATA LINKAGE

Andrew Powierski[1], Rosemary Karmel[2], Phil Anderson[3]

[1] Australian Institute of Health and Welfare
26 Thynne St, Fern Hill Park, Bruce ACT, 2601
andrew.powierski@aihw.gov.au

[2] Australian Institute of Health and Welfare
26 Thynne St, Fern Hill Park, Bruce ACT, 2601
rosemary.karmel@aihw.gov.au

[3] Australian Institute of Health and Welfare
26 Thynne St, Fern Hill Park, Bruce ACT, 2601
phil.anderson@aihw.gov.au

Background:
In Australia, many community service program data collections developed over the last decade, including several for aged care programs, contain a statistical linkage key (SLK) to enable derivation of client level data. In 2005, the Pathways in Aged Care (PIAC) study was funded to create a linked aged care database to enable analysis of pathways through aged care services. The purpose of this paper is twofold:
1. to describe the SLK strategy used to create the PIAC linked database allowing for variation in reported SLK data
2. to compare the results with those using a name-based linkage on the same data.

Methods:
The PIAC linked database was created using a stepwise deterministic record linkage algorithm to link datasets. The strategy uses a general person identifier (an SLK) in conjunction with additional variables. Measures of likely match accuracy were used to select match keys and ensure match quality.
For two large data sets in the PIAC study full name information was available. A name-based linkage strategy was also applied to measure the accuracy of the SLK linkage strategy. The name-based strategy involved running a series of passes allowing for variation in demographic data. After each pass a significant amount of time was spent conducting a clerical review to identify matches where there was variation in reported match data.

Results:
Using the name-based linkage as the reference standard, the SLK strategy had a sensitivity of 98.6% and a positive predictive value of 99.6%. These results confirm the utility of the SLK strategy.
In addition, given the limited match data available, there were many cases in the clerical review where the match status was not clear. That is, name-based linkage strategies themselves are not guaranteed to be 100% accurate.

References
Karmel, Rosemary et al.: Empirical aspects of record linkage across multiple data sets using statistical linkage keys: the experience of the PIAC cohort study. BMC Health Services Research 2010 10:41.
Karmel, Rosemary and Rosman, Diane: Event-based record linkage in health and aged care services data: a methodological innovation. BMC Health Services Research 2007 7:154.

Andrew Powierski *currently works as a Data Analyst with the Australian Institute of Health and Welfare, in Canberra. Andrew has been involved in the analysis of the PIAC linked database. Currently he is completing the study into the comparison of the SLK and name-based linkage strategies.*

IMPROVING ESTIMATED SUFFICIENT SUMMARY PLOTS USING RESIDUAL MINIMIZATION

Luke A. Prendergast

La Trobe University, Bundoora Victoria 3086 Australia
luke.prendergast@latrobe.edu.au

When faced with a large number of regressor variables, dimension reduction methods such as Sliced Inverse Regression (SIR, Li 1991) seek a small number of regressor summary measures that may be used to visually determine a possibly complicated regression structure. Plots of a response versus such summary measures are called Sufficient Summary Plots (SSP's, i.e. Cook 1998) which have proven to be very useful across a diverse range of applications. In this talk we show that estimated SSP's (ESSP's) can sometimes be greatly improved when the dimension reduction estimates are adjusted according to minimization of an objective function. The dimension reduction methods considered include Ordinary Least Squares, SIR and Sliced Average Variance Estimates (SAVE, Cook & Weisberg 1991). The effectiveness of this approach will be highlighted via application to some well known data sets with an emphasis on two- and three-dimensional ESSP's.

References

Cook, R. D. and Weisberg, S. (1991). Discussion of "Sliced Inverse Regression for dimension reduction". Journal of the American Statistical Association, 86, 328-332.
Li, K.-C. (1991). Sliced Inverse Regression for dimension reduction (with discussion). Journal of the American Statistical Association, 86, 316-342.
Cook, R. D. (1998). Regression graphics: Ideas for studying regressions through graphics. New York: John Wiley & Sons Inc..

Luke Prendergast *is a Senior Lecturer in the Department of Mathematics and Statistics at La Trobe University. Luke's research interests largely involve the analysis and visualization of high dimensional data, in particular with respect to dimension reduction methods. As well as teaching undergraduate statistics to students from a diverse range of disciplines, Luke is also heavily involved in the supervision of postgraduate statistics students.*

ADAPTIVE BAYESIAN CLINICAL TRIALS IN CONSERVATION BIOLOGY

William J M Probert[1], Peter W J Baxter[2], Hugh P Possingham[3]

[1] The Department of Mathematics & the Centre for Applied Environmental Decision Making
The University of Queensland, St Lucia, QLD, 4072, Australia
willprobert@gmail.com

[2] Centre for Applied Environmental Decision Making
School of Biological Sciences, The University of Queensland, St Lucia, QLD, 4072, Australia
p.baxter@uq.edu.au

[3] The Department of Mathematics & the Centre for Applied Environmental Decision Making
School of Biological Sciences, The University of Queensland, St Lucia, QLD, 4072, Australia
h.possingham@uq.edu.au

We draw a range of methods from the clinical trial and decision theory literature and apply them to the allocation of endangered species to conservation sites. We are particularly interested in two-armed bandit problems (two treatments to choose between) over a finite horizon with a binary response (success/fail). We assume that at each decision horizon there will be a constant probablility of a superior treatment being found and the experiment stops. Applying methods from the clinical trial literature to non-human subjects also provides new perspectives on ethical considerations in allocating treatments. We illustrate the methods with an example based on an endangered Australian species, the bilby, and conservation efforts in Western Australia to protect it.

Will Probert *is a final year mathematics PhD student at the University of Queensland and part-time research assistant at the Institute for Social Science Research within the university. His PhD thesis is in the area of decision theory, particularly in its application to management decisions in conservation biology. He completed his honours degree in statistics at the University of Otago, New Zealand, with a dissertation on sequential Bayesian methods (particle filters).*

APPLICATION OF FITTING AND VALIDATING TECHNIQUES IN THE DEVELOPMENT OF STATISTICAL EMULATORS FOR A COMPLEX PLANT GROWTH MODEL

Padmaja Ramankutty[1], Megan H Ryan[2], Roger Lawes[3], Jane Speijers[4], Michael Renton[5]

[1] School of Plant Biology and Future Farm Industries CRC, The University of Western Australia and Department of Agriculture and Food WA
35 Stirling Hwy, Crawley, WA 6009, Australia
ramanp01@student.uwa.edu.au

[2] School of Plant Biology and Future Farm Industries CRC, The University of Western Australia
35 Stirling Hwy, Crawley, WA 6009, Australia
megan.ryan@uwa.edu.au

[3] CSIRO Sustainable Ecosystems
Private Bag 5, Wembley WA 6913, Australia
Roger.Lawes@csiro.au

[4] Department of Agriculture and Food WA
3 Baron-Hay Court, South Perth, WA 6151, Australia
jane.speijers@agric.wa.gov.au

[5] School of Plant Biology and Future Farm Industries CRC, The University of Western Australia and CSIRO Sustainable Ecosystems
35 Stirling Hwy, Crawley, WA 6009, Australia
mrenton@cyllene.uwa.edu.au

In this study we aimed to develop statistical emulators for a complex plant growth simulation model. The goal was to find emulators which not only fitted the data generated from the growth simulation well, but which could also provide reliable predictions of new simulation output. To this end, we compared four types of statistical emulators: simple linear models, piecewise linear models, nonlinear models and cubic splines. These types of emulators varied in their complexity both in the different functional forms to fit the main effects and in the method used to determine model parameters. Two methods of model validation were also compared to ensure that the statistical emulator that was ultimately selected provided a good fit to the initial simulation output data and also gave adequate predictions of future simulation output. We illustrate the various techniques of model development and validation using the Agricultural Production Systems sIMulator (APSIM) plant growth model, and consider the effects of rainfall amount, rainfall frequency, temperature and radiation on APSIM-generated biomass production of the perennial pasture species lucerne (Medicago sativa). We conclude that the best statistical emulator in terms of both model fit and ability to predict lucerne biomass production from APSIM is the cubic spline model in which responses for rainfall amount, radiation and temperature are described by linear trends and smooth departures from these trends as represented by splines.

References

Allen, D. M. (1974) The Relationship between Variable Selection and Data Agumentation and a Method for Prediction. Technometrics, 16, 125-127.
Snee, R. D. (1977) Validation of Regression Models: Methods and Examples. Technometrics, 19, 415-428.
Verbyla, A. P., Cullis, B. R., Kenward, M. G. & Welham, S. J. (1999) The Analysis of Designed Experiments and Longitudinal Data by Using Smoothing Splines. Journal of the Royal Statistical Society. Series C (Applied Statistics), 48, 269-311.

Padmaja Ramankutty *is currently a full time PhD student with the School of Plant Biology at the University of Western Australia in Perth. She also works one day per week in the Biometrics section of the Department of Agriculture and Food WA. Padmaja has worked as a statistician in both New Zealand and Australia for the past 10 years and has performed various types of statistical analyses such as sensory trials, repeated measures trials and statistical modelling projects.*

OPTIMUM STRATIFICATION OF A SKEWED POPULATION USING A WEIBULL DISTRIBUTION WITH AUXILIARY INFORMATION

Khan, M.G.M.[1], Reddy, Karuna[2], Rao, Dinesh[3]

[1]School of Computing, Information and Mathematical Sciences
The University of the South Pacific, Suva, Fiji
khan_mg@usp.ac.fj

[2]Faculty of Science and Technology
The University of Fiji, Lautoka, Fiji
KarunaR@unifiji.ac.fj

[3] School of Computing, Information and Mathematical Sciences
The University of the South Pacific, Suva, Fiji
rao_di@usp.ac.fj

Indisputably, optimum strata boundaries (OSB) could be effectively achieved if the distribution of the survey variable, y, is known. This is not feasible in practice since data on a survey variable of interest is unavailable prior to conducting the survey. In practical situations like this, an auxiliary variable, x, which very closely resembles the survey variable can be used instead. Often this variable is highly correlated with y and can be easily made available at low cost from past experience or prior knowledge in a recent survey/study. If stratification is made based on x, it may lead to substantial gains in precision in the estimates, although it will not be as efficient as the one based on y. However, if the correlation between y and x is high within all strata, the boundary points for both the variables should be nearly the same. In such a situation the OSB of the survey variable could be obtained using the frequency distribution of the auxiliary variable. The frequency distribution of the auxiliary variable can be estimated easily as the auxiliary data are readily available or can be made available.

In this manuscript the problem of finding the OSB using an auxiliary variable for a skewed population with a Weibull distribution is considered. The problem is formulated as a mathematical programming problem (MPP) that seeks minimization of the variance of the estimated population parameter under Neyman allocation. The MPP, being a multistage decision problem, is solved using a dynamic programming technique. A numerical example with a real dataset is presented to illustrate the application and computational details of the proposed method. The results are compared with the Dalenius and Hodges' $cum\sqrt{f}$ method, which reveals that the proposed technique is more efficient and also useful for a skewed population.

Karuna Reddy *currently works as an academic at the University of Fiji, in Lautoka City, Saweni Campus, Fiji Islands. His area of interest is Survey Sampling, predominantly, Stratified Random sampling, Mathematical Programming Problems and Operations Research. The main part of his research is determining optimum stratum boundaries (OSB) in Stratified Random Sampling for different types of real-life distributions. Karuna has been involved in teaching, research and consultation in his area of work. Besides this, his interests are Employment Relations, Human Resources Development and Martial Arts Education.*

MAXED OUT? SPECIES DISTRIBUTION MODELLING WITH POISSON POINT PROCESS MODELS

David Warton[1] and Ian Renner[2]

[1] School of Mathematics and Statistics, University of New South Wales
University of New South Wales UNSW NSW 2052 AUSTRALIA
David.Warton@unsw.edu.au

[2] School of Mathematics and Statistics, University of New South Wales
University of New South Wales UNSW NSW 2052 AUSTRALIA
Ian.Renner@unsw.edu.au

Modelling the distribution of a species is a task undertaken by many people in many different disciplines. Consequently, a number of different methods have been developed with this goal in mind, to varying degrees of success. We propose that the best method is to construct a Poisson point process model due to its strong statistical foundation, its flexibility and its equivalence to standard methods of species distribution modelling in other disciplines. One such method known as maximum entropy (or MAXENT) has emerged in the ecology literature and has proven to be very competitive in a comprehensive study of a wide array of species distribution modelling methods. Nevertheless, limitations in the method still exist, ranging from computation time to interpretability of the model output to implementation of smoothing penalties such as the LASSO. We show that a MAXENT model is equivalent to a Poisson point process model evaluated on a regular grid, and re-expressing a MAXENT model as a Poisson Point Process Model addresses many of these limitations. Point process models also offer insight into the ideal spatial resolution at which to evaluate the data for species distribution models, which has been a fundamental issue of species distribution modelling. In light of this, Poisson point process models not only improve and expand leading methods of species distribution models, but offer a new standardization of the methods of implementation.

***Ian Renner** is a 2nd year PhD student at the University of New South Wales, originally from the United States, whose primary interests lie in methodology and education. He has lectured a number of mathematics and statistics courses across three universities in the United States and Australia.*

A SMOOTH TEST OF GOODNESS OF FIT FOR GENERALIZED LINEAR MODELS

Paul Rippon[1], *J.C.W. Rayner*[2]

[1] *University of Newcastle*
University Drive Callaghan NSW 2308 Australia
paul.rippon@newcastle.edu.au

[2] *University of Newcastle*
University Drive Callaghan NSW 2308 Australia
john.rayner@newcastle.edu.au

The approach described by Rayner et al (2009) has been used to derive a smooth test of goodness of fit for generalized linear models. This is applied to several specific GLM examples and the power compared to other common tests. The components which make up the smooth test statistic can also be considered as test statistics in their own right and can provide additional diagnostic information about the lack of fit.

References

Rayner, J.C.W., Thas, O., and Best, D.J. (2009). Smooth Tests of Goodness of Fit: Using R. Wiley.

Paul Rippon *currently works as a Lecturer in Statistics at the University of Newcastle. Originally trained as a Chemical Engineer, Paul became interested in statistics from analysing data to improve industrial processes in line with Total Quality Management principles. While retaining an interest in TQM and industrial statistics, Paul's interests have broadened to include statistical modelling in general as well as statistical education.*

MODELLING THE TRANSMISSION OF HEPATITIS C OVER SOCIAL NETWORKS

G. Daraganova[1], M. Hellard[2], E. McBryde[3,] P. E. Pattison[4], G. L. Robins[5], D. A. Rolls[6]

[1] Department of Psychology
University of Melbourne, VIC 3010
gda@unimelb.edu.au

[2] Centre for Population Health, Burnet Institute
85 Commercial Rd., Melbourne, VIC 3004
hellard@burnet.edu.au

[3] Department of Medicine – Royal Melbourne Hospital
University of Melbourne, VIC 3010
Emma.McBryde@mh.org.au

[4] Department of Psychology
University of Melbourne, VIC 3010
pepatt@unimelb.edu.au

[5] Department of Psychology
University of Melbourne, VIC 3010
garrylr@unimelb.edu.au

[6] Department of Medicine – Royal Melbourne Hospital
University of Melbourne, VIC 3010
drolls@unimelb.edu.au

The hepatitis C virus (HCV) is a blood-borne virus affecting over 170 million people worldwide. In Australia, an estimated 16,000 new infections occur each year. The vast majority of new infections are attributed to injecting drug use. We present results from modelling the transmission of HCV through a network of injecting drug users (IDUs). This work combines empirically grounded models of both the network of IDUs and the transmission of HCV- the latter made possible by an ongoing longitudinal study of the practices and network partners of IDUs. This work is part of a larger project investigating the transmission of infectious diseases through social networks.

Plausible community level network models are made possible by two recent advances. Exponential Random Graph Models with social circuit dependence have shown sufficient flexibility to capture realistic patterns of social contact in a number of applications. New conditional estimation techniques are allowing estimation of community-level network parameters from sample network data. The ability to generate/simulate plausible community-level networks facilitates direct comparisons with other standard assumptions (e.g. homogeneous mixing) and allows investigation of network-specific transmission properties and public health control strategies.

David Rolls *has been a Research Fellow at the University of Melbourne since 2007. He received his Ph.D. from Queen's University at Kingston, Canada in 2003 and worked in the United States before coming to Australia. His research interests generally involve stochastic modelling and simulation. His Ph.D. thesis and research immediately after involved models for data network traffic exhibiting long-range dependence and heavy-tailed distributions. Currently his research is in two areas: the modelling of the transmission of infectious diseases over social networks, and the modelling of financial data with multifractal embedded branching processes.*

HOUSEHOLD-LEVEL INFERENCE FOR EMERGING INFECTIONS

Ross, Joshua V.[1], *House, Thomas*[2], *Keeling, Matt J*[3]

[1] *School of Mathematical Sciences, The University of Adelaide*
joshua.ross@adelaide.edu.au

[2] *Mathematics Institute, University of Warwick*
t.a.house@warwick.ac.uk

[3] *Mathematics Institute and Department of Biological Sciences, University of Warwick;*
m.j.keeling@warwick.ac.uk

During the initial stages of an infection outbreak it is common for authorities to attempt to collect detailed data, corresponding to tracing infections to their source infectious individual. Here we exploit recent work on efficient methods for calculating the distribution of secondary households infected from an individual infected in a household of specified size (Ross et al. (2010)) to calibrate households models to simulated data of the form collected in practice. We consider the case of perfect detection and also the case in which the observations are subject to partial observability; the latter is likely to be commonplace in practice. We pay particular attention to the relationship between accuracy of parameter estimates and the amount of data collected. Finally, we apply our methodology to a data set corresponding to swine (H1N1) 'flu in the U.K. in 2009.

References

Ross, J.V., House, T. and Keeling, M.J. (2010). Calculation of disease dynamics in a population of households. PLoS ONE, 5(3): e9666.

Joshua Ross *is a Lecturer in Applied Mathematics at the University of Adelaide. He was awarded a PhD from the University of Queensland in 2007 and subsequently undertook a one year postdoctoral position at the University of Warwick before a Junior Research Fellowship at King's College, University of Cambridge. His interests are in mathematical and statistical modelling with applications primarily in ecology and epidemiology.*

MODEL ESTIMATION AND TESTS OF HYPOTHESES FOR DIRECTIONAL DATA

Brett Presnell[1], *Pavlina Rumcheva*[2]

[1] *University of Florida, Department of Statistics, Gainesville, FL 32611, U.S.A.*
presnell@stat.ufl.edu

[2] *University of Sydney, School of Public Health, NSW 2006, Australia*
pavlina.rumcheva@sydney.edu.au

We consider the spherically projected multivariate linear model for directional data in the general d-dimensional case. Maximum likelihood estimates for the model are readily computed using iterative methods. We develop multi-sample likelihood ratio tests for equality of mean directions and concentrations. The performance of the tests in terms of size and power is demonstrated with computer simulation. A closed-form expression for the mean resultant length of the d-dimensional projected normal distribution is presented. An example involving anthropological data illustrates the application of the methods in three dimensions.

References

Presnell, B. and Rumcheva, P. (2008). The mean resultant length of the spherically projected normal distribution. Statistics & Probability Letters, 78, 557-563.

Presnell, B., Morrison, S.P., and Littell, R.C. (1998). Projected multivariate linear models for directional data. Journal of the American Statistical Association, 93, 1068-1077.

Pavlina Rumcheva *currently works as a Lecturer at the School of Public Health, University of Sydney, in Australia. Her areas of interest include directional data and survival analysis. She is presently involved in postgraduate teaching and public health research.*

DESIGNS FOR EXPERIMENTS WHERE OBSERVATIONS ARE TAKEN OVER TIME

K.G. Russell

University of Wollongong
Centre for Statistical and Sampling Methodology and School of Mathematics & Applied Statistics,
University of Wollongong NSW 2522
kgr@uow.edu.au

We consider designed experiments where observations are taken repeatedly on each experimental unit. While considerable knowledge has been gained about how to analyse the data from such experiments, little theory has yet been developed on the optimal design of the experiment.

It is desired to determine
• at what times observations should be taken, and
• at what levels of the between-units factor(s)
observations should be made in order to optimise the experiment.

Examples of such experiments include industrial investigations where each batch is the result of a combination of factors, and the researchers wish to examine properties of the batch over time.

The principles underlying the determination of these optimal designs will be described, and examples of such designs will be given.

This is joint work with Professor Susan Lewis and Dr David Woods of the University of Southampton.

Ken Russell *is currently an Associate Professor in Statistics with the Centre for Statistical and Survey Methodology at the University of Wollongong. His area of research is Experimental Design.*

FORECASTING ABUNDANCE OF KEY SPECIES FROM THE COMMERCIAL FISHERY IN PORT PHILLIP BAY, VICTORIA

Karina L Ryan[1], Denny Meyer[2] and Khageswor Giri[3]

[1] Fisheries Victoria, Department of Primary Industries
PO Box 114, Queenscliff, Victoria 3225, Australia
karina.ryan@dpi.vic.gov.au

[2] Faculty of Life and Social Sciences, Swinburne University of Technology
PO Box 218, Hawthorn, Victoria, 3122, Australia
dmeyer@swin.edu.au

[3] Future Farming Systems Research Division, Department of Primary Industries
600 Sneydes Road, Werribee, Victoria, 3030, Australia
khageswor.giri@dpi.vic.gov.au

Predictive models that forecast fisheries abundance can assist routine stock assessments. Autoregressive Integrated Moving Average (ARIMA) models reveal trend and seasonality patterns and may assist in developing an 'adaptive' management framework. 'Adaptive' management frameworks can allow for less stringent management controls when abundance is high (and environmental conditions are most suitable) or tightening controls when abundance is low (and environmental conditions are least suitable). Absolute estimates of stock abundance are difficult and expensive to obtain, but catches by commercial fishers can provide a regular "sample" of the population. An index of relative abundance is obtained from the catch per unit effort (CPUE) by dividing total catch by the effort required to obtain that catch. Predictive models that forecast CPUE can potentially utilise explanatory variables that relate to environmental patterns (e.g. zonal westerly winds; Jenkins 2005) or stock structure (e.g. size/age of individual cohorts and an index of pre-recruits from fishery independent surveys; Stockhausen and Fogarty 2007). Time series models were constructed using CPUE of key species from commercial fishing in Port Phillip Bay during a 30 year period from 1978/79 to 2008/09. The efficacy of univariate and multivariate models were compared for short-lived species (southern calamary), for those species that are represented by 1-2 years classes (King George whiting) and by 5-6 year classes (snapper). For King George whiting, an ARIMA$(1,0,2)(0,1,1)_{12}$ was the best model for forecasting purposes when compared with an auto regression using an environmental variable. The ARIMA model had independent residuals, the lowest Schwarz Bayesian Criterion and no non-significant coefficients. ARIMA models are most suitable for providing short-term forecasts (up to 12 months). Although multivariate models might be more appropriate for determining the effects of additional explanatory variables, they may not provide better forecasts and are dependent on an appropriate time series of explanatory variables.

References

Jenkins GP (2005). The influence of climate on the fishery recruitment of a temperate, seagrass associated fish, the King George whiting, Sillaginodes punctata. Marine Ecology Progress Series 288:263-271

Stockhausen WT, Fogarty MJ (2007). Removing observational noise from fisheries–independent time series data using ARIMA models. Fishery Bulletin 105(1):88–101.

Karina Ryan *has worked previously with the Queensland Department of Primary Industries, University of Western Australia and Victorian Fisheries Research Institute with a focus on indigenous, recreational and commercial fishing. She joined the Department of Primary Industries in 2004 to evaluate methods of obtaining total catch estimates for individual Victorian bay and inlet recreational fisheries and conduct a large-scale survey of recreational fishing using an angler licence sampling frame. Karina currently oversees the finfish fisheries and stock assessments for Fisheries Victoria. She is a member of the Australia Society for Fish Biology and the Statistical Society of Australia and is currently completing a Master of Science (Applied Statistics) with Swinburne University. This paper presents results from the research component of this degree.*

OUTBREAK DYNAMICS OF AN ENDEMIC HORTICULTURAL PEST OF EASTERN AUSTRALIA

Sadler. R. J.[1], B. White[2], V. Florec[3], B. C. Dominiak[4]

[1] CRC Plant Biosecurity, The University of Western Australia
School of Agricultural and Resource Economics 35 Stirling Highway, Crawley, WA 600
rsadler@cyllene.uwa.edu.au

[2] CRC Plant Biosecurity, The University of Western Australia
School of Agricultural and Resource Economics, 35 Stirling Highway, Crawley, WA 6009
benedict.white@uwa.edu.au

[3] CRC Plant Biosecurity, The University of Western Australia
School of Agricultural and Resource Economics, 35 Stirling Highway, Crawley, WA 6009
veronique.florec@uwa.edu.au

[4] CRC Plant Biosecurity
Plant Biosecurity Risk Management, Industry & Investment
161 Kite Street, Orange, NSW 2800
bernie.dominiak@industry.nsw.gov.au

The Queensland fruit fly (Qfly; Bactrocera tryoni) is an endemic pest of eastern Australian horticulture, putting at risk a multi-billion dollar industry. Area wide management of Qfly, which includes surveillance, quarantining and eradication, is a public good: the cost is borne by the taxpayer whereas the benefits are largely realised by private enterprise. To maintain significant public expenditure in support of area wide measures, such as pest free status, requires Benefit-Cost Analyses (BCA) that quantify the risk of failure of different management strategies and technologies. This risk however is driven by the ecological dynamics of Qfly outbreaks and human transport of infested produce. We provide a spatio-temporal model of Qfly outbreaks for incorporation into a BCA. Using weekly monitoring data collected at some 3000 sites over 10 years, we estimate parameters using a method of inference based on Diggle & Gratton's (1984) Monte Carlo simulation of the likelihood surface. We amend Diggle & Gratton's parameter search mechanism with a particle swarm. More broadly statistically implicit models, where a mathematical expression for the likelihood is either difficult or unable to be written down, are common in ecology theory. Inference for such models is necessary if ecology is to inform the current shift towards evidence-based policy for agricultural and environmental landscapes.

References

Diggle P.J. and R.J. Gratton (1984). Monte Carlo methods of inference for implicit statistical models (with discussion). Journal of the Royal Statistical Society Series B 46:193-227.

Rohan Sadler *is a Research Assistant Professor employed by the School of Agricultural and Resource Economics at the University of Western Australia (UWA). Rohan completed his PhD at UWA in 2007, and has since worked on areas of agricultural and environmental policy, with an emphasis on incorporating remote sensed information into models of ecological dynamics that underpin policy decision tools. A key area of research has also been in inferring fire regimes, and the interplay between prescribed fire and wild fire.*

SIMULATION OF COAL-BARGING ON A RIVER

Hana Sakai[1], *Emma Smith*[2]

[1] *Data Analysis Australia*
97 Broadway, Nedlands WA 6009
hana@daa.com.au

[2] *Data Analysis Australia*
97 Broadway, Nedlands WA 6009
emma@daa.com.au

A coal mining company was intending on a river of variable navigability according to rainfall to barge coal from a barge loading terminal to a transhipment location in the ocean. Data Analysis Australia was contracted to determine the feasibility of such a process, and to estimate the limits (in tones per year) that could be transported. A simulation program was developed, using the Extend software package, and sensitivities to certain elements such as the number of barge passing-points, ship arrival rates and reliability, and the number of barges running were investigated. The simulation took into account loaded and unloaded barge speeds, priorities at passing points, river availability and maximum barge queue lengths.

Hana Sakai *has been working as a Consultant Statistician at Data Analysis Australia since December 2008. Having completed a degree in Business Decision Analysis in the UK, Hana was previously working as an Operational Research Analyst for the British Government. Whilst at Data Analysis Australia, a selection of Hana's work includes predicting throughput for the mining industry, analysing the results of various surveys to determine pricing strategies and modelling the effect of daylight saving on water consumption.*

MULTIPLATFORM SMOOTH SEGMENTATION FOR THE IMPROVED DETECTION OF COPY NUMBER VARIATIONS

Agus Salim[1], Shu Mei Teo[2], Yudi Pawitan[3]

[1] Department of Epidemiology and Public Health, Yong Loo Lin School of Medicine
National University of Singapore
Block MD3 Level 3, 16 Medical Drive, 117597, Singapore
ephaguss@nus.edu.sg

[2] Department of Epidemiology and Public Health, Yong Loo Lin School of Medicine
National University of Singapore
Block MD3 Level 3, 16 Medical Drive, 117597, Singapore
g0801862@nus.edu.sg

[3] Department of Medical Epidemiology and Biostatistics, Karolinska Institutet
Nobels vag 12A, Stockholm 17177, Sweden
yudi.pawitan@ki.se

With the rapid expansion of whole genome studies, there is rapid evolution of genotyping platforms. This leads to practical issues such as constant upgrading of genotyping equipment. In the context of DNA copy number estimation, it would mean that the same research group could potentially estimate copy number of the same samples, using from different platforms. While having more data could potentially eventuate in more precise and accurate copy number estimates, combining such data is not straightforward because it is known that estimates from different platforms show different degrees of attenuation of the true copy number changes (Bengtsson et al., 2009). Furthermore, different platforms have different noise characteristics and different marker panels. An ideal solution would be to make joint copy number calling using data from different platforms simultaneously. Recently, Zhang et al (2010) proposed a multiplatform circular binary segmentation (MPCBS) algorithm for joint estimation of copy number from multiple platforms. Huang et al (2007) showed that the circular binary segmentation method does not work optimally when there is considerable contamination of normal cells in tumor samples. Here, we propose an algorithm for joint copy number calling based on an extension of the Huang et al (2007) method for single platform. The method is based on correlated random-effects for the unobserved copy number pattern. We compare our method to MPCBS using normal and tumor samples and we show that our method performs as well as, and occasionally better than MPCBS in detecting copy number regions, especially in tumor samples.

References

Bengtsson, H et al., 2009. A single-sample method for normalizing and combining full-resolution copy numbers from multiple platforms, labs and analysis methods. Bioinformatics 25(7), pp. 861-867

Huang, J et al., 2007. Robust smooth segmentation approach for array CGH data analysis. Bioinformatics 23(18), pp.2463-2469.

Zhang, NR., Senbabaoglu, Y., Li, JZ. 2010. Joint estimation of DNA copy number from multiple platforms. Bioinformatics 26(2), pp. 153-160.

Agus Salim *obtained his PhD in Statistics from National University of Ireland in 2003. Subsequently, he spent almost four years at the Australian National University as a postdoctoral fellow, mainly working on developing methods for novel epidemiological study designs such as case-series, case-crossover and nested case-control data. In 2008, he moved to the National University of Singapore and since then has also been interested in statistical applications in molecular biology and genetics.*

FITTING LINEAR MIXED MODELS USING PROBABILISTICALLY LINKED DATA

Klairung Samart

Centre for Statistical and Survey Methodology
University of Wollongong, Wollongong NSW 2522
ks208@uow.edu.au

Probabilistic matching of records from different data sets is often used to create linked datasets for use in research in health, epidemiology, economics, demography and sociology. Clearly, this type of matching can lead to linkage errors, which in turn can lead to bias and increased variability when standard statistical estimation techniques are used with the linked data. Chambers (2009) describes an inferential framework for statistical modelling using probabilistically linked data, which is then used to develop modified estimation methods for regression models based on the assumption that the correctly linked data are mutually uncorrelated. In real life, however, measurements are usually made on clusters of correlated statistical units, such as people in a family, patients in a hospital or students in a school, and when analysing such data, linear mixed models are often used. In this paper we show how the inferential framework of Chambers (2009) can be used to develop unbiased regression parameter estimates when fitting a linear mixed model to probabilistically linked data. Furthermore, since estimation of variance components is also an important objective when fitting a mixed model, we develop appropriate modifications to standard methods of variance components estimation in order to account for linkage error. In particular, we focus on three widely used methods of variance components estimation: ANOVA, maximum likelihood (ML) and restricted maximum likelihood (REML) (Searle, Casella & McCulloch, 2006). Modified versions of all three methods that allow for linkage errors are described, and simulation results comparing them are presented.

References:

Chambers, R. (2009). Regression analysis of probability-linked data. Statisphere, Official Statistics Research Series, Volume 4.

Searle, S. R., Casella, G., and McCulloch, C. E. (2006). Variance Components. New York: John Wiley & Sons.

Klairung Samart *received her B.Sc. (Honours) in Mathematics from Prince of Songkla University, Thailand in 2005 and her Grad.Dip (Teaching) from Chiang Mai University, Thailand in 2006. She was a lecturer in the Department of Mathematics, Faculty of Science, Prince of Songkla University in 2006. Following an award of a scholarship from Prince of Songkla University to do a Masters in Australia, she received her M.Stat. (with Merit) from the University of Newcastle in 2008. She is currently pursuing her PhD in Statistics at the University of Wollongong under a University Postgraduate Award (UPA) from the University of Wollongong and a scholarship from Prince of Songkla University.*

ESTIMATION OF REGRESSION PARAMETERS WHEN MEASUREMENT ERROR IN EXPLANATORY VARIABLE

Saqr A[1], Khan S[2]

[1] Department of Mathematics & Computing
Australian Centre for Sustainable Catchments
University of Southern Queensland, QLD, 4350, Australia
and Department of Statistics, AlJabal AlGarbiya University, Libya
anw_sag@yahoo.com

[2] Department of Mathematics & Computing
Australian Centre for Sustainable Catchments
University of Southern Queensland, QLD, 4350, Australia
khans@usq.edu.au

This paper proposes a new estimation method for the parameters of a simple linear regression model where the explanatory variable is subject to measurement error. The new estimator of the regression parameters is based on the reflection of the available values of the explanatory variable. The method is straightforward, easy to implement, and performs much better than existing estimators. Moreover, it does not depend on any unrealistic assumptions about the value of any scale parameter or the reliability ratio or independence of measurement error and model error. The theoretical superiority of the proposed estimator is established by analytical results. An illustrative example for numerical comparisons is also included.

Anwar Saqr *is a PhD candidate at the University of Southern Queensland, QLD, 4350, Australia and a lecturer (on leave) from the Department of Statistics, AlJabal AlGarbiya University, Libya. His main area of interest is measurement error in linear and nonlinear models. Anwar is now working on instrumental variable and reflection methods to deal with measurement errors.*

NUMERICAL EVALUATION OF THE CDF AND QUANTILE FUNCTIONS FOR THE GENERALIZED HYPERBOLIC AND RELATED DISTRIBUTIONS

David J Scott[1], Joyce Li[2], Thanh Tam Tran[3]

[1] Department of Statistics, The University of Auckland, PB 92019, Auckland New Zealand, d.scott@auckland.ac.nz

[2] Department of Statistics, The University of Auckland, PB 92019, Auckland New Zealand, xli053@aucklanduni.ac.nz

[3] Department of Statistics, The University of Auckland, PB 92019, Auckland New Zealand, thtam@stat.auckland.ac.nz

The generalized hyperbolic distribution is a very flexible and diverse family which can exhibit extreme skewness and peakedness. It also has (semi)-heavy tails. These properties make the writing of fast accurate routines for the cumulative distribution function and quantile function quite challenging. We describe the best approaches we have currently found for this problem.

References

Scott, David J. (2009). GeneralizedHyperbolic: The generalized hyperbolic distribution. R package version 0.2-0. Accessed from http://CRAN.R-project.org/package=GeneralizedHyperbolic

David Scott *is Associate Professor of Statistics at the University of Auckland. He has authored a number of R packages which are available on CRAN implementing functions for distributions. His main research interest is in statistical computing. Besides teaching and research he is an experienced statistical consultant who advises clients both from within the University of Auckland and from outside.*

ENVIRONMENTAL CONVERGENCE AND SUSTAINABILITY: THE CASE OF SOUTHEAST ASIA

Siok Kun Sek[1], Wai Mun Har[2]

[1] School of Mathematical Sciences, Universiti Sains Malaysia
11800 Minden, Penang, Malaysia
sksek@usm.my

[2] Faculty of Accountancy and Management, Universiti Tunku Abdul Rahman
Bandar Sungai Long, 43000 Selangor, Malaysia
harwm@mail.utar.edu.my

The impact of climate change and the effect of greenhouse have received increasing attention from the policy makers and public. In particular, emissions of carbon dioxide are of concern globally. Scientists find that the continual increase in carbon dioxide (CO2) emissions and other greenhouse gases has contributed to higher global temperatures (Lee & Chang, 2009). Focusing the analysis on several Southeast Asian countries, we seek to investigate the convergence of carbon dioxide emissions on the sustainability of environmental quality in Southeast Asian countries. The convergence of carbon dioxide implies sustainability of environmental quality and the allocation of carbon dioxide emissions without resource transfers through international trade and cross-border movements of industries (Romero-Ávila, 2008). The unit-root tests are applied to detect the behavior of shocks, (permanent or temporary) and the stationarity of the series. Stationarity implies relative emissions of carbon dioxide are temporary and there is a proof of stochastic convergence. The unit-root or non-stationarity reveals that shocks are permanent and there is no convergence in the relative emissions of carbon dioxide. The results reveal evidences on the convergence of carbon dioxide emissions in East-Asia.

References

Lee C.C & Chang, C.P. (2009). Stochastic convergence of per capita carbon dioxide emissions and multiple structural breaks in OECD countries. Economic Modelling 26: 1375-1381.
Romero-Ávila, D. (2008). Convergence in carbon dioxide emissions among industrialised countries revisited. Energy Economics 30: 2265-2282.

Siok Kun Sek *currently works as a senior lecturer in the School of Mathematical Sciences, Universiti Sains Malaysia. She obtained her doctoral degree from Christian-Albrechts University of Kiel, Germany. Her area of interest is econometrics, International Macroeconomics and Finance.*

NONPARAMETRIC MODELLING AND FORECASTING OF ELECTRICITY DEMAND: AN EMPIRICAL STUDY

Han Lin Shang

Department of Econometrics & Business Statistics, Monash University
Building H, 900 Dandenong Road, Caulfield East, Victoria 3800, Australia
HanLin.Shang@buseco.monash.edu.au

This paper uses half-hourly electricity demand data in South Australia as an empirical study of nonparametric modelling and forecasting methods for predictions from a half-hour ahead to one year ahead. A notable feature of the univariate time series of electricity demand is the presence of both intraweek and intraday seasonalities. An intraday seasonal cycle is apparent from the similarity of the demand from one day to the next, and an intraweek seasonal cycle is evident from comparing the demand on the corresponding day of adjacent weeks. There is a strong appeal in using forecasting methods that are able to capture both seasonalities. In this paper, the forecasting methods slice a seasonal univariate time series into a time series of curves. The forecasting methods reduce the dimensionality by applying functional principal component analysis to the observed data, and then utilize a univariate time series forecasting method and functional principal component regression techniques. When data points in the most recent curve are partially observed, the updating methods can improve the point forecast accuracy. We also revisit a nonparametric approach to construct prediction intervals of updated forecasts, and evaluate the interval forecast accuracy.

Han Lin Shang *currently works as a research fellow at the Department of Econometrics & Business Statistics, Monash University, Caulfield campus, in Melbourne. His area of interest is functional data analysis, multivariate data analysis and demographic forecasting. He has published articles in the Journal of Computational and Graphical Statistics, the Journal of the Korean Statistical Society and Mathematics and Computers in Simulation.*

SPATIAL POINT PATTERN ANALYSIS OF BREAST CANCER MORTALITY IN PERTH, 2005

C. Shao[1], U. Mueller[2], J. Cross[3]

[1] School of Engineering, Edith Cowan University
100 Joondalup Drive, Joondalup, WA, 6027, Australia
c.shao@.ecu.edu.au

[2] School of Engineering, Edith Cowan University
100 Joondalup Drive, Joondalup, WA, 6027, Australia
u.mueller@ecu.edu.au

[3] School of Engineering, Edith Cowan University
100 Joondalup Drive, Joondalup, WA, 6027, Australia
j.cross@.ecu.edu.au

Breast cancer mortality data for the Perth metropolitan area are analysed using spatial point process models. The data sets comprise breast cancer incidence and mortality data recorded by the Department of Health of Western Australia (WA) during the period 1990-2005. The individual data records consist of the following variables: Latitude/longitude of location at incidence (death); age in 5-year brackets at time of diagnosis (death) and the year at time of diagnosis (death).

Breast cancer incidence and mortality will be considered for females only. Covariates available for the mortality data are female population density, age of the female cancer patients at the time of diagnosis and spatial intensity of the vulnerable population, modelled by cancer incidence data for Perth. Gaussian Kernel smoothing is used to smooth the female population density and the proportion of females aged 40 or above derived from the year 2006 population data by suburb. Covariates including the incidence of other cancers and spatial trends are explored to model the breast cancer mortality point patterns. The approach is exemplified using the 2005 data. Both the breast cancer incidence and mortality patterns can be modelled as inhomogeneous Poisson processes. The fitted models will be validated via lurking variable plots and plots of the residuals derivative against covariate. The quality of the fit is shown to be dependent on the nature of the covariates. The results show that the female population density is a more suitable covariate than the age of cancer patients at the time of diagnosis.

Changying Shao *is currently a PhD student in Edith Cowan University, under the supervision of Associate Professor Ute Mueller and Associate Professor James Cross. Her project is modelling cancer data using geostatistical and spatial methods.*

STATISTICAL POWER CALCULATION AND SAMPLE SIZE DETERMINATION FOR ENVIRONMENTAL STUDIES WITH DATA BELOW DETECTION LIMITS

Quanxi Shao[1], *You-Gan Wang*[2]

[1] *CSIRO Mathematics, Informatics and Statistics*
Leeuwin Centre, 65 Brockway Road, Floreat, WA 6014, Australia
Quanxi.Shao@csiro.au

[2] *Centre for Applications in Natural Resource Management*
School of Mathematics and Physics, The University of Queensland
St Lucia, Brisbane 4072, Australia

Power calculation and sample size determination are critical in designing environmental monitoring programs. The traditional approach based on comparing the mean values may become statistically inappropriate and even invalid when substantial proportions of the response values are below the detection limits or censored because strong distributional assumptions have to be made on the censored observations when implementing the traditional procedures. By noting that environmental studies are frequently interested in percentiles instead of mean values, we propose a quantile methodology that is robust to outliers and can also handle data with a substantial proportion of below-detection-limit observations without the need of imputing the censored values. As a demonstration, we applied the methods to a nutrient monitoring project, which is a part of the Perth Long-Term Ocean Outlet Monitoring Program. In this example, the sample size required by our quantile methodology is, in fact, smaller than that by the traditional t-test, illustrating the merit of our method.

References

Quanxi Shao and You-Gan Wang (2009). Sample size determination for environmental studies when data may fall below detection limits. Water Resources Research 45, W09410. doi:10.1029/2008WR007563..

Quanxi Shao *is currently a principal research scientist in CSIRO Mathematics, Informatics and Statistics. His area of interest is nonparametric inference, model/variable selection, quantile approach, time series analysis and spatial statistics. He has worked in a multidisciplinary environment and managed research teams in several national projects, focusing on statistical applications to environmental study including hydrology and water resources research.*

A STUDY ON CLUSTER ANALYSIS FOR ORDER-RESTRICTED DATA WITH APPLICATION TO EPISODIC MEMORY DATA ANALYSIS

Tokiko Shimizu[1], Toru Ogura[2], Toshinari Kamakura[3]

[1] Graduate School of Science and Engineering, Chuo University
1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan
kagachiin.cou@gmail.com

[2] Chuo University
1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan
ogura@indsys.chuo-u.ac.jp

[3] Chuo University
1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan
kamakura@indsys.chuo-u.ac.jp

Cluster analysis for order-restricted data plays an important role in the fields of healthcare and ubiquitous technologies (Matthew & Anind, 2008). For instance, episodic memory data clustering is a very important task for detecting symptoms of dementia and anepia patients, and observations are arranged on a time scale. Duration of time from memorizing is one of the most important factors that must be considered.

Cluster analysis sometimes may detect a group of patients with characteristic features that we can use for diagnosis. However, standard-clustering methods cannot handle order-restricted variables. Changing the order of variables does not give rise to different results, whereas we would expect a different result according to changed orderings. In this article, we propose a new statistical clustering method for order-restricted data based on order restriction techniques with a pool adjacent violator algorithm.

Firstly, we apply the proposed method to real multiple time point episodic memory data. Secondly, we simulate various order-restricted data and illustrate that our proposed method works well.

References

Matthew L. Lee & Anind K. Dey (2008). Lifelogging Memory Appliance for People with Episodic Memory Impairment, UbiComp, 344, 44-53.

Tokiko Shimizu *is currently a Graduate Student in the Graduate School of Science and Engineering, Chuo University, Kourakuen Campus, in Tokyo. Her area of interest is Cluster Analysis for Order-restricted Data.*

ON THE ESTIMATION OF POPULATION MEAN IN TWO-STAGE SAMPLING UNDER NON-RESPONSE

V. K. Singh[1], M. K. Chaudhary[2]

[1]Department of Statistics
Banaras Hindu University
Varanasi – 221005 INDIA
vijay_usha_2000@yahoo.com

[2]Department of Statistics
Banaras Hindu University
Varanasi – 221005 INDIA
ritamanoj15@gmail.com

In real world situations the population to be surveyed generally has different configurations in relation to elementary and sampling units. As per objective of the survey, therefore, it is sometimes not feasible or it is administratively inconvenient to select elementary units directly from the population using the simple random sampling. Multi-stage sampling has been found to be very useful in such situations. Moreover, it is now widely recognized that non-sampling errors, particularly non-response errors, need as much attention as sampling errors in the design and execution of sample surveys. There is now considerable literature on the treatment of non-response error. Different types of techniques and methods for assessing and controlling non-response error have been developed. The method of sub-sampling of non-respondents, present in the sample, is comparatively simple and widely acceptable.

The present paper deals with the development of a family of estimators for estimating population mean in two–stage sampling with equal size clusters under the presumption that the population is subject to non-response error. The family exhibits some good properties and consists of some well-known estimators as special cases. The results have been illustrated with the help of some empirical data.

V K Singh *received a Master's Degree and Ph. D. Degree in Statistics in 1972 and 1979 respectively from Banaras Hindu University. He joined the Department of Statistics, Banaras Hindu University, Varanasi as a Lecturer in 1972 and became Professor of Statistics in 1998. He has 38 years of teaching experience in undergraduate and post graduate programmes. Branches of specialization are Sampling Theory, Stochastic Processes, Operations Research and Demography. He has published about 65 research papers in reputed journals, guided 12 research scholars in Sampling Theory for their Ph. D. Degree and completed two research projects. He is associate Editor of the Assam Statistical Review journal and a referee for a number of Indian Statistical journals. He visited Brighton, U.K. in 1988 and Melbourne, Australia in 1997 to deliver invited lectures at International Conferences. He has worked as Head, Department of Statistics, Banaras Hindu University, Varanasi from August, 2007 to July, 2010 And is presently, sharing the responsibility of Deputy Coordinator, Special Assistance Programme (DRS – I) of the Department.*

LONGITUDINAL CA125 ALGORITHM (ROCA) FOR EARLY DETECTION OF OVARIAN CANCER

[1]Steven J. Skates, [2]Usha Menon, [3]Ian J. Jacobs

[1] Massachusetts General Hospital and Harvard Medical School
50 Staniford Street, Suite 560, Boston MA 02114, USA
sskates@partners.org

[2] Gynaecological Cancer Research Center, University College London
UCL Elizabeth Garrett Institute for Women's Health, Maple House
149 Tottenham Court Road, London W1T 7DN, UK
u.menon@ucl.ac.uk

[3] Faculty of Biomedicine, University College London
Faculty Offices, 149 Tottenham Court Road, London W1T 7JA, UK
i.jacobs@ucl.ac.uk

Early detection of ovarian cancer is an appealing approach to reducing mortality from this disease for which mortality has been constant for over four decades. Most cases (80%) are detected in late stage disease with a very poor prognosis, while if detected in early stage prognosis is excellent. The challenge to early detection is the relatively low incidence with one case annually in 2,500 postmenopausal women. CA125 blood test is a known ovarian cancer biomarker with a standard positive result when the test exceeds 30IU/mL. However, sensitivity for early stage disease only increases to 35-40%. Longitudinal analysis of past screening trials indicates each woman has her own CA125 baseline level, and that significant increases above the baseline flag women with high probability of disease. We developed a algorithm based on statistical modelling of longitudinal CA125 values (Skates et al, 2003), which in simulations and pilot studies increased the sensitivity for early stage disease to 60% at the same specificity level. We implemented it in a 5 year pilot study of 14,000 women (Menon et al, 2005), refined it, and then implemented the new version in a 15 year definitive randomized study of 200,000 women with ovarian cancer mortality as the endpoint. The development and implementation of this risk of ovarian cancer algorithm (ROCA), and promising results from the prevalence screen (Menon et al, 2009) will be described.

References

Skates SJ, Menon U, MacDonald N, Rosenthal AN, Oram DH, Knapp RC, Jacobs IJ. Calculation of the risk of ovarian cancer from serial CA-125 values for preclinical detection in postmenopausal women. J Clin Oncol. 2003 May 15;21(10 Suppl):206s-210s.

Menon U, Skates SJ, Lewis S, Rosenthal AN, Rufford B, Sibley K, Macdonald N, Dawnay A, Jeyarajah A, Bast RC Jr, Oram D, Jacobs IJ. Prospective study using the risk of ovarian cancer algorithm to screen for ovarian cancer. J Clin Oncol. 2005 Nov 1;23(31):7919-26.

Menon U, Gentry-Maharaj A, Hallett R, Ryan A, Burnell M, Sharma A, Lewis S, Davies S, Philpott S, Lopes A, Godfrey K, Oram D, Herod J,Williamson K, Seif MW, Scott I, Mould T, Woolas R, Murdoch J, Dobbs S, Amso NN, Leeson S, Cruickshank D, McGuire A, Campbell S,Fallowfield L, Singh N, Dawnay A, Skates SJ, Parmar M, Jacobs I. Sensitivity and specificity of multimodal and ultrasound screening for ovarian cancer, and stage distribution of detected cancers: results of the prevalence screen of the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS). Lancet Oncol. 2009 Apr;10(4):327-40.

Steven J. Skates *currently works as a cancer researcher at the Massachusetts General Hospital Cancer Center in Boston, MA. His area of interest is the early detection of ovarian cancer, including development of screening algorithms, conducting screening trials using these algorithms, and searching for new ovarian cancer biomarkers.*

LATTICE PATH APPROACH FOR TRANSIENT ANALYSIS OF M/G/1 QUEUES UNDER (0,k) CONTROL POLICIES USING C_2 COXIAN DISTRIBUTION

Isnandar Slamet[1], Ritu Gupta[2], and Narasimaha Achuthan[3]

[1]Department of Mathematics, Sebelas Maret University
Jl Ir Sutami 36 A Kentingan Surakarta 57126 Indonesia
isnandar06@yahoo.com

[2]Department of Mathematics and Statistics, Curtin University of Technology
GPO BOX U1987 Perth 6845 Western Australia
R.Gupta@curtin.edu.au

[3]Department of Mathematics and Statistics, Curtin University of Technology
GPO BOX U1987 Perth 6845 Western Australia
N.R.Achuthan@curtin.edu.au

The queuing systems with server vacations have interesting applications in optimally scheduling the server time. The transient solutions for such systems are required to optimize the server usage. In this paper we present transient analysis of a M/G/1 queueing process that operates under (0,k) vacation policy, wherein the server goes on vacation when the system becomes empty and re-opens for service immediately at the arrival of the kth customer.

The transient analysis is based on approximating the general service time distribution by a Coxian two-phase distribution and representing the corresponding queueing process as a lattice path. Finally lattice path combinatorics are used to present the transient solution. The key advantage of the approach is simplicity of mathematical structure, explicit solutions and numerical computations. The simulation of (0,k) systems will be presented to illustrate these advantages.

References

Borkakaty, B., Agarwal, M. and Sen, K. (2010). Lattice path approach for busy period density of GIa /Gb/1 queues using C2 Coxian distributions. Applied Mathematical Modelling. 34 (6): p. 1597-1614.
Muto, K., Miyazaki, H., Seki, Y. Kimura, Y., and Shibata, Y. (1995). Lattice path counting and M/M/c queueing systems. Queueing Systems. 19: p. 193-214.
Cox, D.R., (1955). A use of complex probabilities in the theory of stochastic processes. Mathematical Proceedings of the Cambridge Philosophical Society, 51: p. 313-319.

Isnandar Slamet *is a PhD student at the Department of Mathematics and Statistics, Curtin University of Technology, WA. His area of interest includes statistical inference, statistical computing, time series and stochastic modelling. Currently he is doing research in the area of queueing systems.*

SPATIAL MODELS WITH COMPLEX CONTIGUITY MATRICES IN SMALL AREA ESTIMATION

G. Yu. Sofronov

Department of Statistics, Macquarie University NSW 2109 Australia
georgy.sofronov@mq.edu.au

Estimation of population characteristics for sub-national domains (or smaller regions) is an important objective for statistical surveys. In particular, geographically defined domains, e.g. regions, states, counties, wards and metropolitan areas can be of interest. One of the popular methods in small area estimation (SAE) is the use of linear mixed models with area specific random effects to account for between areas variation beyond that explained by auxiliary variables included in the fixed part of the model (see Rao (2003)). In order to use spatial auxiliary information in SAE, it is reasonable to assume that either the area or the individual random effects (defined, for example, by a contiguity criterion) are correlated, with the correlation decaying to zero as the distance between these areas increases (Pfeffermann, 2002). We study the effect of using different contiguity matrices on SAE. Estimation of the mean squared error of the resulting small area estimators is discussed. The properties of the estimators are evaluated by applying them to the results of farm surveys that have been conducted by the Australian Bureau of Agricultural and Resource Economics.

References

Pfeffermann, D. (2002). Small area estimation - new developments and directions. International Statistical Review, 70, 125-143.
Rao, J. N. K. (2003). Small Area Estimation. New York: Wiley.

Georgy Sofronov *gained his Ph.D. in Statistics in 2002 at Moscow State University. He has held research positions at the University of Queensland and the University of Wollongong. Currently he is a Lecturer in Statistics at Macquarie University. His research interests include Markov Chain Monte Carlo simulation, small area estimation, the Cross-Entropy method, change-point problem and optimal stopping rules.*

A LARGE-SCALE GENOME SIMULATION MODEL INCORPORATING PATTERNS OF LINKAGE DISEQUILIBRIUM AND OTHER POPULATION GENETIC PARAMETERS FOR USE IN THE AUSTRALIAN DAIRY CATTLE POPULATION

Jie Song[1,2], *Mehar Khatkar*[2], *Herman Raadsma*[2], *Peter Thomson*[2]

[1] *Prince of Wales Clinical School, University of New South Wales, Kensington, NSW*

[2] *ReproGen – Animal Bioscience Group, Faculty of Veterinary Science*
University of Sydney, 425 Werombi Road, Camden, NSW
json7944@uni.sydney.edu.au

For genome wide association (GWA) studies, a large number of hypotheses are tested simultaneously, with each hypothesis being the existence or otherwise of an association between the trait and the SNP or haplotype. This multiple testing problem can be tackled by controlling False Discovery Rate (FDR) (Benjamini and Hochberg, 1995). The correlation between the tests will affect the estimation of FDR. If the correlation is high, it will be difficult to discern which of a number of tightly linked SNPs might be directly associated with the trait of interest, and this will result in high FDR. The Linkage Disequilibrium (LD) is one factor that influences the correlation between the tests, so the interest here is to investigate how LD affects the reliable identification of true SNP-trait associations based on the Australian dairy cattle population, and this is undertaken by means of a simulation study. The aim of this simulation study is to make a simulated population match as closely as possible to the LD structure of the Australian Holstein-Friesian dairy cattle population.

The model adopted forward simulation of the population and included parameters for mutation, varied recombination rates along the chromosome, migration, base gamete population size and distribution of minor allele frequency of SNPs in the base gamete population. The best simulation scenario with the closest LD pattern was found by combining the different parameter settings. The mean square error (MSE) was used here as the criterion for determining which simulation scenario is the best fitting one to the Australian dairy cattle data. The best simulation scenario was the one with the smallest MSE.

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological), 57, 289-300.

Jie Song *just finished her PhD study with the Faculty of Veterinary Science, University of Sydney and currently works as a research associate with Prince of Wales Clinic School, University of New South Wales. Her area of interest is the statistical methodology development for genomic data.*

BAYESIAN CONTROL CHART MONITORING

Elizabeth Stojanovski

Lecturer, School of Mathematical and Physical Sciences, University of Newcastle
Elizabeth.Stojanovski@newcastle.edu.au

The implementation of a Bayesian statistical process control chart is presented. The model assumes the control level to have a prior probability distribution as opposed to only a fixed value for the purpose of providing parameter values for the monitored process. The posterior distribution of the relevant parameter is consequently calculated. The detection of a shift is based on information within the posterior. A quality control case study is presented to identify how different interventions have influenced the rate of smoking care provision within a clinical setting over time. Change in clinical practice behaviour with regard to cessation care was evident across the study period, with the greatest change being noted in response to combinations of interventions. Here, the variable of interest is the fraction defective or the fraction of smokers who were provided with cessation advice during their clinical visit. The analysis concentrates on Bayesian control charts for monitoring the process fraction.

Elizabeth Stojanovski *currently works as a Statistics Lecturer in the School of Mathematical and Physical Sciences at the University of Newcastle. Her area of interest is in applied statistics, with a focus on Bayesian methodology and meta-analyses. Elizabeth has been involved extensively in consultation and research with staff from the Faculty of Health.*

A SINGLE-INDEX MODEL FOR DOWNSCALING IN A SPATIAL SETTING

Alex Stuckey

Australian Bureau of Statistics
alex.stuckey@abs.gov.au

A statistical downscaling model is proposed using a single-index model with index coefficients forming a smooth surface over a spatial grid. An application to modelling decreasing rainfall in south-west Western Australia is given.

The single-index model is a well understood semiparametric method popular in many fields, in particular econometrics. Whereas multivariate nonparametric modelling can fall victim to the curse of dimensionality, the high-dimensional explanatory variable is reduced to a single-index by means of a linear combination. The coefficients of this index are estimated for optimal prediction of the response variable where the link function is estimated nonparametrically (Härdle, Hall and Ichimura, 1993).

In a statistical downscaling problem motivated by climatic studies, researchers may wish to model local (eg. south-west Western Australia) behaviour of one variable (rainfall) as a function of behaviour of another variable (mean sea-level pressure) measured over a much larger region (the southern hemisphere)(Li and Smith, 2009).

In this context a single-index model can be developed where the index is a linear combination of the pressure measurement at each site. As the number of pressure sites and therefore the number of index coefficients to be estimated can be very large, a further dimension reduction measure is proposed. This measure is to constrain the index coefficients to be defined by a member of a class of smooth surfaces. This surface can be defined by means of orthogonal polynomials, including spherical harmonics. Methods for fitting such a model are explored, as are applications.

References

Härdle, W., Hall, P. and Ichimura, H. (1993). Optimal smoothing in single-index models. Annals of Statistics ,21, 157-178.
Li, Y. and Smith, I. (2009). A Statistical downscaling model for southern Australia winter rainfall. Journal of Climate ,22, 1142-1158.

Alex Stuckey *works in the Time Series Analysis section of the Australian Bureau of Statistics. This paper describes his PhD done within the School of Mathematics and Statistics of the University of Western Australia under the supervision of Prof. Jiti Gao (University of Adelaide), Dr Yun Li (CSIRO Mathematical and Information Sciences) and Assoc. Prof. Robin Milne (University of Western Australia).* *The study was funded by UWA and CSIRO.*

A NEW INFLUENCE MEASURE FOR PRINCIPAL COMPONENTS WITH APPLICATIONS TO HIGH DIMENSIONAL DATA SETS

Luke A. Prendergast[1],*Connie Li Wai Suen*[2]

[1] *La Trobe University, Victoria 3086 Australia*
luke.prendergast@latrobe.edu.au

[2] *La Trobe University, Victoria 3086 Australia*
c.liwaisuen@latrobe.edu.au

Principal Component Analysis (PCA) is an important tool in multivariate analysis, in particular when faced with high dimensional data. There has been much done with regards to sensitivity analysis and the development of influence diagnostics for the eigenvector estimators that define the sample principal components. However, little, if any, has been done in this setting with regards to the sample principal components themselves. In this paper we develop a sensitivity measure for principal components associated with the covariance matrix that is very much related to the influence function (Hampel, 1974). This influence measure is based on the average squared canonical correlation and differs from existing measures in that it assesses influence of certain observational types on the sample principal components. We use this measure to derive an influence diagnostic that satisfies two key criteria:(i) it detects influential observations with respect to the retained sample principal components and (ii) it is efficient to calculate even in high dimensions. We use microarray data sets to show that our measure satisfies both criteria.

References

Hampel, F. R. (1974). The influence curve and its role in robust estimation. Journal of the American Statistical Association, 69, 383-393.

Connie Li Wai Suen *is a PhD student at La Trobe University, Melbourne. Her area of interest is robust statistics for high dimensional data as well as applications in meta analysis. Connie has been working as a casual tutor for the past few years and is also heavily involved in teaching R to undergraduate students.*

USING SOCIAL NETWORK INFORMATION IN SURVEY ESTIMATION

Thomas Suesse[1], *Raymond Chambers*[2]

[1] *Centre for Statistical and Survey Methodology, University of Wollongong*
Wollongong, NSW 2522, Australia
tsuesse@uow.edu.au

[2] *Centre for Statistical and Survey Methodology, University of Wollongong*
Wollongong, NSW 2522, Australia
ray@uow.edu.au

Standard statistical approaches for social statistics obtained from sample surveys, censuses and administrative sources usually focus on individuals. For example, a survey may be used to collect information on smoking, drinking and exercise behaviour as well as and other social and demographic variables from a sample of individuals. Naive single level analysis of these data that assumes individuals behave independently is usually inappropriate and can lead to incorrect inferences. A more appropriate approach is a multilevel one, with levels corresponding to households and geographic areas to account for correlations among members of the same household or area.

Here we focus on the use of social networks as another source of information when modelling survey variables. For example, best friends form a social network which can be assumed to significantly influence individual behaviour. Standard modelling approaches for this type of social influence are based on spatial autocorrelation and disturbance models, where the weight matrix reflects the structure of the social network. Here we propose an alternative approach that treats the social network information as another level in a multi-level model. We also develop contextual models, where an individual's characteristics are assumed to depend on the average characteristics of other individuals in his/her network.

In this paper we consider the effect of the inclusion of such social network information on estimation of population totals, focusing in particular on the case where the social network information is only collected from sampled individuals. In a simulation study we investigate and compare the efficiency of the BLUP and GREG based on four different social network models. In doing so, we aim to address a fundamental question: Is it worthwhile to collect social network information for use in sample survey estimation?

Thomas Suesse *currently works as a postgraduate research fellow at the Centre of Statistical and Survey Methodology of the University of Wollongong. He works closely with Profs David Steel and Ray Chambers in an ARC/ESRC Linkage International project on the role of households, neighbourhoods and networks in social statistics. The project also involves collaboration with Dr Mark Tranmer from the University of Manchester, UK. He obtained his PhD degree in statistics from Victoria University of Wellington, New Zealand. His research focuses on survey methodology, social networks and categorical data analysis.*

ON EXPECTED LENGTH OF THE IMPROVED PREDICTION INTERVALS FOR INDEPENDENT OBSERVATIONS

Khreshna Syuhada[1], *Udjianna Pasaribu*[2], *Utriweni Mukhaiyar*[3]

[1] *Statistics Research Division-Institut Teknologi Bandung (ITB), Jalan Ganesa 10 Bandung, khreshna@math.itb.ac.id*

[2] *Statistics Research Division-Institut Teknologi Bandung (ITB), Jalan Ganesa 10 Bandung, udjianna@math.itb.ac.id*

[3] *Statistics Research Division-Institut Teknologi Bandung (ITB), Jalan Ganesa 10 Bandung, utriweni@math.itb.ac.id*

Assessment of the accuracy of prediction interval via expected length is derived. Prediction intervals we consider are the estimative and improved prediction intervals for future independent observations. It is found that we can not assess the estimative prediction interval through its expected length since there can be a trade off between the $O(n^{-1})$ term in the asymptotic expansion of the coverage probability and in the asymptotic expansion of the expected length. Furthermore, we show that the bias of the parameter estimator has a significant effect on the expected length of the prediction interval.

References

Barndorff-Nielsen, O.E., Cox, D.R. (1994). Inference and Asymptotics. London: Chapman and Hall.
Kabaila, P., Syuhada, K. (2007). The relative efficiency of prediction intervals. Communications in Statistics: Theory and Methods, 36(15), 2673-2686.
Kabaila, P., Syuhada, K. (2008). Improved prediction limits for AR(p) and ARCH(p) processes. Journal of Time Series Analysis 29(2), 213-223.

Khreshna Syuhada *completed a PhD Degree at La Trobe University last year. Khreshna works as a Lecturer at Institut Teknologi Bandung (ITB), Indonesia. His research areas include financial time series, Value-at-Risk, and volatility modelling. Khreshna has FOUR children.*

MIXED MODEL VARIABLE SELECTION (MMVS): AN APPLICATION OF QTL ANALYSIS WITH COMPLEX MIXED MODELS

Julian Taylor[1], Ari Verbyla[2]

[1] Mathematics, Informatics and Statistics, CSIRO, Adelaide, Australia
julian.taylor@csiro.au

[2] Mathematics, Informatics and Statistics, CSIRO, Adelaide, Australia and School of Agriculture Food and Wine, University of Adelaide, Adelaide, Australia
ari.verbyla@csiro.au

One common focus in modern plant breeding experiments is the analysis of Quantitative Trait Loci (QTL). Unfortunately these experiments are often complex requiring non-genetic sources of variation to be captured such as structured spatial correlations between observations and/or variation arising from other experimental design components. This talk will show a whole genome analysis of QTL can be achieved using a variable selection method which is succinctly incorporated into current mixed model theory. The methodology presented can also be used when the number of genetic markers is greater than the number of observations making it an ideal tool for high dimensional analysis. To ensure efficiency the Mixed Model Variable Selection (MMVS) method uses the flexible software package ASReml-R (Butler, et al. 2009) as its core linear mixed model fitting routine. Under simulation the MMVS method shows that the power of QTL detection is increased for varying population sizes in comparison to the forward selection procedure, Whole Genome Average Interval Mapping (WGAIM) presented in Verbyla et al. (2007). This talk will also discuss the use of MMVS as a tool for discovering higher order epistatic interactions in complex designed plant breeding experiments.

References

Butler, D. G and Cullis, B. R and Gilmour, A. R and Gogel, B. J (2009). ASReml-R Reference Manual. Queensland Department of Primary Industries.

Verbyla, A. P and Culli, B. R and Thompson, R. (2007). The analysis of QTL by simultaneous use of the full linkage map. Theoretical and Applied Genetics, 116, 95-111.

Julian Taylor *is currently employed at the Mathematics, Informatics and Statistics division of CSIRO based in Adelaide where he is affiliated with the Food Futures National Research Flagship. He currently collaborates on projects with Plant Industry including the analysis of QTL for wheat quality with advanced intercrosses. His skills include statistical modelling including linear mixed models, asymptotics and computational statistics.*

SPATIAL ASSESSMENT OF CHARTER FISHING IN THE WEST COAST BIOREGION OF WESTERN AUSTRALIA

Carli Telfer[1], Ute Mueller[2], Brent Wise[3], Glenn Hyndes[4]

[1] Department of Fisheries and Edith Cowan University
Po Box 20 North Beach WA 6920
carli.telfer@fish.wa.gov.au or ctelfer@our.ecu.edu.au

[2] Edith Cowan University
270 Joondalup Drive, Joondalup WA 6027
u.mueller@ecu.edu.au

[3]Department of Fisheries
Po Box 20 North Beach WA 6920
brent.wise@fish.wa.gov.au

[4]Edith Cowan University
270 Joondalup Drive, Joondalup WA 6027
g.hyndes@ecu.edu.au

In Western Australia, the Tour Operators and Aquatic Eco-Tourism (Charter) sector forms a part of the fishing community. Management of the sector aims to ensure its long-term sustainability together with the natural resources it depends on. The aim of this study was to assess the charter sector's catch rate data between 2002/03 and 2007/08 at a bioregional level using various geostatistical techniques to determine how the spatial structure of catch rate information has changed over time. Knowledge of the spatial and temporal distribution of fishing characteristics is essential when trying to better understand what impact a sector may have on fish populations and the environment.

Experimental semivariograms were constructed to assess the spatial continuity of the charter sector catch rate data for three West Coast bioregion key indicator species, Pagrus auratus, Choerodon rubescens and Glaucosoma hebraicum over a six (6) year period. The results show an increase in the ranges of semivariogram models for C. rubescens catch rates, while for G. hebraicum and P. auratus there was variability in the ranges but no consistent increase or decrease. Geostatistical analysis also highlighted changes in local catch rate densities within the bioregion, completely independent to the species. Overall the charter sector in the West Coast bioregion has demonstrated moderate change in the spatial behaviour over the study period. This was particularly evident in 2005/06 and 2006/07 when the range of variability dropped from 219 nautical miles to 20 nautical miles within two years; further investigation showed that this change occurred around the Perth metropolitan area perhaps as a result of operators searching for fish stocks and/or targeting particular species. Ordinary Kriging estimates showed that the majority of low catch rate estimates also occurred near Perth; this was notably evident in the catch rates for *P.* auratus.

Carli Telfer *currently works as a Research officer at the Department of Fisheries, Research Division in Perth Western Australia. Her area of interest is in the spatial analysis and interpretation of recreational fisheries data using geostatistical techniques. Carli first developed these skills and interest whilst undertaking research for her Master's degree she recently completed through Edith Cowan University in 2010.*

INCREASING THE PRECISION OF GENETIC ASSOCIATION ANALYSES THROUGH MULTIPLE IMPUTATION OF MISSING GENOTYPES

Lidija Turkovic[1], John B. Carlin[1 2], Lyle C. Gurrin[1]

[1]Centre for Molecular, Environmental, Genetic & Analytic Epidemiology
Melbourne School of Population Health 1/723 Swanston Street
Carlton, Victoria, 3010, Australia
lidija78@yahoo.com
lgurrin@unimelb.edu.au

[2]Murdoch Childrens Research Institute
Royal Children's Hospital
Parkville, Victoria 3052, Australia
john.carlin@mcri.edu.au

Imputation of genotypes offers the opportunity to explore disease associations at untyped loci by exploiting the linkage disequilibrium (correlation) between SNP genotypes observed in a reference panel. While the accuracy of genotype imputation has been investigated extensively, relatively little has been done on examining accuracy of estimates of regression parameters in association analysis. We explore this in a simulation study by using samples from 865 people to genotype SNPs in 12 genes of iron metabolism (Constantine et al 2009). We simulated binary outcomes by using target SNPs of different minor allele frequencies (MAFs) and a range of odds ratios in an additive genetic model. We then took sub-samples from the generated data, set a random proportion of the participants to have missing data at some locations and performed genotype imputation. The association between SNPs and outcome was then examined using logistic regression. The precision and bias of estimates were evaluated for different sets of parameter values and proportions of missing data with imputation using IMPUTE (Marchini et al 2007), Beagle (Browining & Browning 2007) and Stata (ice). Analysis has been completed for all common SNPs typed in *CYBRD1* gene using IMPUTE. Results show that for an odds ratio of 2.00 estimates of log odds ratios show only small bias. There is a big improvement in precision after imputation of missing genotypes over analysis that is restricted to complete cases with genotype data at all loci. For SNPs with high (0.39), moderate (0.09) and low (0.02) MAF, the variance ratio for parameter estimates comparing a partial dataset with all loci typed to a full dataset after imputation were 8.22, 5.20 and 1.65 respectively. For SNPs with moderate to high MAF multiple imputation offers a substantial gain in precision of parameter estimates.

References

Browning, S. and Browning, B. (2007). Rapid and accurate haplotype phasing and missing data inference for whole genome association studies using localized haplotype clustering. *American Journal of human Genetics*, 81, 1084-1097.

Constantine, C.C., Anderson, G.J., Vulpe, C.D., McLaren, C.E., et al. (2009). A novel association between a SNP in CYBRD1 and serum ferritin levels in a cohort study of HFE hereditary haemochromatosis. *British Journal of Haemotology*, 147, 140-149.

Marchini, J., Howie, B., Myers, S., McVean, G. and Donnelly, P. (2007). A new multipoint method for genome-wide association studies via imputation of genotypes. *Nature Genetics*, 39, 906-913.

Lidija Turkovic *has a Bachelors degree in Mathematics and Computer Science and a Masters degree in Statstics. She is currently completing a PhD at the Centre for Molecular, Environmental, Genetic and Analytic Epidemiology, University of Melbourne. Her area of interest is imputation of missing genotypes in genetic association studies. She is also an investigator of the NHMRC-funded HealthIron study, examining genetic and enviromental modifiers of hereditary haemochromatosis (iron overload disease) which is part of the Melbourne Collaborative Cohort Study.*

SEMIPARAMETRIC REGRESSION WITH SHAPE CONSTRAINED P-SPLINES

Martin L Hazelton[1], Berwin A Turlach[2]

[1] Institute of Fundamental Sciences - Statistics, Massey University
Manawatu PN461, Private Bag 11222, Palmerston North, New Zealand
m.hazelton@massey.ac.nz

[2] School of Mathematics and Statistics (M019), University of Western Australia
35 Stirling Highway, CrawleyWA 6009, Australia
berwin@maths.uwa.edu.au

In semiparametric regression models, P-splines can be used to describe complex, non-linear relationships between the mean response and covariates. In some applications it is necessary, or at least desirable, to restrict the shape of the splines so as to enforce properties such as monotonicity or convexity on regression functions. We describe a method for imposing such shape constraints on P-splines within a linear mixed model framework. We employ MCMC methods for model fitting, using a truncated prior distribution to impose the requisite shape restrictions. We develop a computationally efficient MCMC sampler by using a corrrespondingly truncated multivariate normal proposal distribution, which is a restricted version of the approximate sampling distribution of the model parameters in an unconstrained version of the model. We also describe a cheap approximation to this methodology that can be applied for shape constrained scatterplot smoothing. We illustrate our methods on an example involving monotonic growth curves for trees under different experimental conditions.

Berwin A Turlach *is Associate Professor in the School of Mathematics and Statistics at the University of Western Australia. He received a degree as Diplom-Mathematiker from the University of Bonn, Germany, in 1991, and his Diplôme d'Etudes Approfondies en mathématique and Docteur en Statistique from the Université Catholique de Louvain in Louvain-la-Neuve, Belgium, in 1992 and 1994, respectively. From January 1995 to June 1998 he worked at the Australian National University. The first two years as Research Associate at the Centre for Mathematics and Its Applications and then as Research Fellow at the Cooperative Research Centre for Advanced Computational Systems. From June 1998 to December 1999 he was Lecturer at the Department of Statistics of the University of Adelaide, South Australia. Since December 1999, he has worked at the University of Western Australia, Western Australia, with a break from December 2006 to June 2009 during which he worked as Associate Professor at the Department of Statistics and Applied Probability of the National University of Singapore. His research interests include nonparametric smoothing methods, computational statistics and applied statistics.*

STATISTICAL PROCESS CONTROL FOR MONITORING LUNG FUNCTION IN ASTHMA

Robin M Turner[1], Andrew Hayen[2], Petra Macaskill[3], Les Irwig[4], Helen K Reddel[5]

[1] Screening and Test Evaluation Program, Sydney School of Public Health
Edward Ford Building (A27), The University of Sydney, NSW 2006, Australia
robin.turner@sydney.edu.au

[2] Screening and Test Evaluation Program, Sydney School of Public Health
Edward Ford Building (A27), The University of Sydney, NSW 2006, Australia
andrew.hayen@sydney.edu.au

[3] Screening and Test Evaluation Program, Sydney School of Public Health
Edward Ford Building (A27), The University of Sydney, NSW 2006, Australia
petram@health.usyd.edu.au

[4] Screening and Test Evaluation Program, Sydney School of Public Health
Edward Ford Building (A27), The University of Sydney, NSW 2006, Australia
lesi@health.usyd.edu.au

[5] Woolcock Institute of Medical Research and University of Sydney
431 Glebe Point Road, Glebe, NSW 2037
hkr@med.usyd.edu.au

Statistical process control charts have been proposed for use in monitoring of lung function in asthma to rapidly detect future changes, particularly exacerbations. Control limits, based on the variability of the process, are used to detect when the process moves out of control. The limits need to be set during a period of stability. Previous studies of control charts in patients with asthma have generally reported case studies. To assess their wider application, we have used Shewhart X-mR charts on existing data from a randomised trial of 80 patients with asthma to assess the statistical control of the lung function measurements and to detect a treatment change. Patients were on maximal therapy at the beginning of the trial and had optimal lung function, measured using peak expiratory flow (PEF) and forced expiratory volume in 1 second (FEV1). Control charts were created for each patient using the last 20 days prior to randomisation to treatment change to set the control limits. The proportion of patients in statistical control during this period was assessed using 5 control chart rules. Despite being on optimal therapy the number of patients with PEF measurements in statistical control ranged from 59% to 78% for different combinations of the rules, with similar proportions for FEV1. Unexpectedly for patients in statistical control prior to randomisation, there was no difference in alerts on PEF measurements (p=0.7) in patients randomised to change treatments (8% signalled a decrease) than in patients continuing the same treatment (11%). However, for FEV1 35% of the patients randomised to treatment change had a decrease compared to 6% for patients continuing the same treatment (p=0.004). The lack of statistical control for patients on maximal therapy makes the use of control charts in the routine monitoring of lung function problematic in many patients.

Robin Turner *currently works as a Research Fellow (Biostatistics) in the Screening and Test Evaluation Program, School of Public Health at the University of Sydney. Her areas of interest are screening and diagnostic tests, disease monitoring, and the use of statistical methods in epidemiology.*

CONDITIONAL CONFIDENCE OF BINOMIAL INTERVALS

Frank Tuyl[1], Richard Gerlach[2], Kerrie Mengersen[3]

[1] School of Mathematical and Physical Sciences, University of Newcastle
Callaghan 2308 NSW, Australia
frank.tuyl@newcastle.edu.au

[2] Faculty of Economics and Business, University of Sydney
Sydney 2006 NSW, Australia
richard.gerlach@sydney.edu.au

[3] School of Mathematical Sciences, Queensland University of Technology
Brisbane 4001 QLD, Australia
k.mengersen@qut.edu.au

Conditional confidence of a given interval is considered in the face of possible frequentist objection. When applied to intervals for the Normal mean, for example, there is no difference between frequentist confidence and Bayesian probability, but for the binomial parameter there are many different methods of interval calculation. Here, frequentist `exact' methods are based on the minimum coverage requirement, while `approximate' methods relax this constraint and focus on mean coverage instead. Bayesian intervals based on noninformative priors may be compared with the latter family; in fact, when based on the uniform or Bayes-Laplace prior, mean coverage is exactly equal to nominal, which is not achieved by any of the usual approximate methods, such as Wald, Score and mid-P.

From a practitioner's point of view, the apparent inadequacy of both exact and approximate methods, when applied to simple examples of given intervals (including ones from Bayes' original paper), throws doubt on their suitability for estimation of the binomial parameter when the sample size is small. In contrast, intervals viewed as having been derived from the Bayes-Laplace posterior, which is simply the normalised likelihood function, are shown to lead to logical and acceptable probability statements.

Frank Tuyl *After many years in industry and health, Frank is currently a lecturer with the University of Newcastle, Australia. Frank is trying to convince frequentists and Bayesians alike that when the aim is to "let the data speak for themselves", the binomial and Poisson parameters are best estimated by highest posterior density (HPD) intervals based on uniform priors.*

AN APPLICATION OF THE SOUND RECOGNITION TECHNIQUE TO THE HUMAN ACTION ACKNOWLEDGMENT

Itsuki Uemizo[1], Kousuke Okusa[2], Toshinari Kamakura[3]

[1] Graduate School of Science and Engineering, Chuo University, 1-13-27 Kasuga, Bunkyoku, Tokyo, i_uemizo@indsys.chuo-u.ac.jp

[2] Graduate School of Science and Engineering, Chuo University, 1-13-27 Kasuga, Bunkyoku, Tokyo, k.okusa@me.com

[3] Department of Science and Engineering, Chuo University, 1-13-27 Kasuga, Bunkyoku, Tokyo, kamakura@indsys.chuo-u.ac,jp

In recent years, researchers in the medical or industrial field have become interested in the statistical analysis of human acts and their recognition for safety and security purposes. The important problem is cause investigation and prevention of the accidents in the medical or reliability application. Kuwahara (2003) suggests the method for preventing accidents based on the analysis of the dataset obtained from acceleration sensors. Takeuchi (2009) studied the method to recognize daily action (standing, sitting, etc) using acceleration sensors.

Previous works mainly used Short Time Fourier Transform and SVM etc. However, these studies can detect very simple activity only (pace of walking, standing, sitting), for observations are infrasonic frequency waves, and such waves are not suitable for standard FFT methods.

In this article, we propose a new method for recognition of human acts using Dynamic Time Warping, which is the technique of the sound recognition. We illustrate that this technique is very useful even for discriminating human acts.

References

Noriaki, Kuwahara., Haruo, Noma., Nobuji, Tetsutani., Norihiro, Hagita., Kiyoshi, Kogure. and Hiroshi, Iseki(2003). Auto-event-recording for Nursing Operations by Using Wearable Sensors. AISJ Journal 2003, 2638-2648.

Shinichi, Takeuchi., Shinya, Itou., Satoshi, Tamura. and Satoru, Hayamizu(2009) Human Activity Recognition Based on Acceleration Information IEICE Technical Report CAS2008-142, SIP2008-205, CS2008-116(2009-3), 229-234.

Itsuki Uemizo *currently works as a Graduate Student with the Statistical Data Analysis Laboratory, Graduate School of Science and Engineering, Chuo University, Kourakuen Campus, in Tokyo. His area of interest is the Pattern Recognition of Human Activity.*

STOCHASTIC MODELLING OF VOLATILITY AND INTER-RELATIONSHIPS IN THE AUSTRALIAN ELECTRICITY MARKETS

Joanna Wang[1] *and Jennifer Chan*[2]

[1] *School of Mathematics and Statistics*
The University of Sydney, Australia
joannaw@maths.usyd.edu.au

[2] *School of Mathematics and Statistics*
The University of Sydney, Australia
jchan@maths.usyd.edu.au

To model price and volatility of the Australian wholesale spot electricity markets, the univariate generalised autoregressive conditional heteroskedasticity (GARCH) models have been applied and the inter-relationships in these markets are modelled using multivariate GARCH models. Stochastic volatility (SV) models, as exible alternatives to GARCH models, have demonstrated their superiority in many _nancial applications. However, the use of SV models in the modelling of electricity markets is still quite limited. This paper investigates existing multivariate SV models and proposes new SV models with skew error distributions, to model the price and volatility of three pairs of markets, selected from four regional electricity markets in Australia, which are shown to be highly correlated in a previous study (Higgs, 2009). Bayesian approach using Markov chain Monte Carlo method is adopted and model implementation is done using the software WinBUGS. Empirical results show that the price and volatilities of selected markets are strongly correlated across di_erent pairs of regional markets. Based on Deviance Criterion Information, the models with skew error distributions perform better than those with symmetric distribution.

Joanna Wang *is currently a PhD student in the School of Mathematics and Statistics, the University of Sydney. Her main research interests focus on stochastic volatilities models using Bayesian computational methods and scale mixtures density representation. She is currently working on the use of heavy-tailed and skewed error distributions for stochastic volatility models.*

NONPARAMETRIC INFERENCE PROCEDURE FOR RANDOM EFFECTS META ANALYSIS

Rui Wang[1], *Lu Tian*[2], *Tianxi Cai*[3], *LJ Wei*[4]

[1] *Massachusetts General Hospital and Harvard School of Public Health*
655 Huntington Ave., Boston, MA 02115, USA
rwang@hsph.harvard.edu

[2] *Department of Health Policy and Research, Stanford University School of Medicine*
Stanford, California 94305, USA
lutian@stanford.edu

[3] *Harvard School of Public Health, 655 Huntington Ave., Boston, MA 02115, USA*
tcai@hsph.harvard.edu

[4] *Harvard School of Public Health, 655 Huntington Ave., Boston, MA 02115, USA*
wei@hsph.harvard.edu

Suppose that the random effects distribution of the parameter of interest in meta analysis is completely unspecified. We propose a nonparametric interval estimation procedure for making inferences about the percentiles, for example, the median, of this distribution. In contrast to the existing methods which can only make inferences about the center of the random effects distribution, the validity of the new proposal does not require the number of studies involved to be large. The new proposal is theoretically valid when the sample sizes of individual studies are large. Empirically we find that our procedure performs well even with moderate individual study sample sizes. The new procedure can be implemented with study-level summary statistics. We apply the method to analyze the data from a recent study to investigate the potential treatment-related toxicity from erythropiesis-stimulating agents (ESAs) for treating cancer patients with anemia (Bennett et al., 2008). In contrast to the conclusions from Bennett et al., our results suggest that the ESAs may be beneficial for certain subgroups of cancer patients with respect to mortality. On the other hand, for cardiovascular toxicity, our results are consistent with those published in the literature. That is, cancer patients treated with ESAs tend to experience more venous thromboembolism events than those untreated.

References

Bennett, C. L. Silver, S. M., Djulbegovic, B., et al. (2008). Venous thromboembolism and mortality associated with recombinant Erythropoietin and Darbepoetin administration for the treatment of cancer-associated anemia. Journal of the American Medical Association, 299, 914-924.

Rui Wang *currently works as a Biostatistician at the Massachusetts General Hospital and Harvard School of Public Health, in Boston, USA. Her areas of interest are design and analysis of clinical trials.*

POISSON POINT PROCESS MODELS SOLVE THE "PSEUDO-ABSENCE PROBLEM" FOR PRESENCE-ONLY DATA IN ECOLOGY

David I. Warton[1], *Leah C. Shepherd*[2]

[1] *The University of New South Wales*
School of Mathematics and Statistics and Evolution & Ecology Research Centre
The University of New South Wales, NSW 2052, Australia
David.Warton@unsw.edu.au

[2] *The University of New South Wales*
School of Mathematics and Statistics and Evolution & Ecology Research Centre
The University of New South Wales, NSW 2052, Australia
Leah.Shepherd@iinet.net.au

Presence-only data, point locations where a species has been recorded as being present, are often used in modelling the distribution of a species as a function of a set of explanatory variables -- whether to map species occurrence, to understand its association with the environment, or to predict its response to environmental change. Currently, ecologists commonly analyze presence-only data by adding randomly chosen "pseudo-absences" to the data such that it can be analyzed using logistic regression, an approach which has weaknesses in model specification, in interpretation, and in implementation. To address these issues, we propose Poisson point process modelling of the intensity of presences. We also derive a link between the proposed approach and logistic regression - specifically, we show that as the number of pseudo-absences increases (in a regular or uniform random arrangement), logistic regression slope parameters and their standard errors converge to those of the corresponding Poisson point process model. We discuss the practical implications of these results. In particular, point process modelling offers a framework for choice of the number and location of pseudo-absences, both of which are currently chosen by ad hoc and sometimes ineffective methods in ecology, a point which we illustrate by example.

References

Warton D.I. & Shepherd L.C. (in press). Poisson point process models solve the "pseudo-absence problem" for presence-only data in ecology. Annals of Applied Statistics.

David Warton *is an ecological statistician based in the Statistics department at UNSW, whose cross-disciplinary research involves evaluating the methods for data analysis currently used in ecology, and where necessary, developing new methodologies to assist ecologists answer key research questions. He has made methodological contributions in multivariate analysis (with applications in allometry), high-dimensional data analysis (with applications in community ecology), and point process modelling (with applications in species distribution modelling). Amongst other things, he is currently interested in combining these last two interests.*

LOCALIZATION ALGORITHM IN THE INDOOR ENVIRONMENT BASED ON THE ToA DATA

Taishi Watabe[1], *Toshinari Kamakura*[2]

[1] *Graduate School of Science and Engineering, Chuo University*
1-13-27 Kasuga,Bunkyo-ku
t_watabe@indsys.chuo-u.ac.jp

[2] *Department of Science and Engineering, Chuo University*
1-13-27 Kasuga,Bunkyo-ku
kamakura@indsys.chuo-u.ac.jp

Recently, a highly accurate localization technique in the indoor environment plays a very important role in the field of ubiquitous society. We propose a new localization algorithm by use of observed ToA (Time-of-Arrival) range data. The main problem of ToA-based range measurements in indoor environments is that it is very difficult to model errors by multipath and NLoS(Non-Line-of-Sight) signal propagation. In this article we propose a new method designed to reduce the effects of NLoS errors using statistical properties of the observed data. When we use the proposed technique the location estimates can be easily calculated even when there are two or more nodes existing in the NLoS environment. We compared the accuracy of the proposed technique and two or more related ones.

References

Rohrig, C. and Muller M. (2009) Localization of Sensor Nodes in a Wireless Sensor Network Using the nanoLOC TRX Transceiver.Proc.VTC Spring 2009.

Taishi Watabe *currently works as a Graduate Student with the Statistical Data Analysis Laboratory, Graduate School of Science and Engineering, Chuo University, Kourakuen Campus, in Tokyo. His area of interest is the Localization of Sensor Nodes in a Wireless Sensor Network.*

THE IMPACT OF INTRODUCING COMPUTER ASSISTED PERSONAL INTERVIEWING TO THE HILDA SURVEY

Nicole Watson

Melbourne Institute of Applied Economic and Social Research, University of Melbourne
Level 7, Alan Gilbert Building, University of Melbourne VIC 3010
n.watson@unimelb.edu.au

The Household, Income and Labour Dynamics in Australia (HILDA) Survey made the shift from pen-and-paper methods to computer assisted personal interviewing (CAPI) in wave 9. The CAPI technology promises significant benefits in data quality, greater in-field control, and improved monitoring of response rates. On the other hand, use of this technology also brings the risk of a break in data continuity.
The issues that will be examined include:

i) how the introduction of CAPI affected the data in terms of item non-response, completeness to multi- and open-ended questions, and continuity in key variables;

ii) what impacts the new technology had for overall respondent burden;

iii) the effects of dependent interviewing (feeding forward data from a prior wave); and

iv) whether these impacts vary across population sub-groups.

Nicole Watson *has worked on the HILDA project at the Melbourne Institute of Applied Economic and Social Research, University of Melbourne, since 2000. She has managed the fieldwork contracts for the first 9 waves and has been involved extensively with the editing, imputation, weighting and data documentation for the survey.*

IMPUTATION OF MISSING DATA IN A SELF-EXCITING MODEL FOR TERRORIST ACTIVITY VIA BAYESIAN IMPUTATION AND PARALLEL LIKELIHOOD COMPUTATION

Gentry White[1], Micheal Porter[2], Lorraine Mazerolle[3]

[1]Australian Research Council Centre of Excellence in Policing and Security
The University of Queensland, Brisbane, Queensland 4072 Australia
gentry.white@uq.edu.au

[2] SPADAC, McLean, VA USA
mike.porter@spadac.com

[3] Australian Research Council Centre of Excellence in Policing and Security
The University of Queensland, Brisbane, Queensland 4072 Australia
l.mazerolle@uq.edu.au

Self-exciting models have been used to describe both temporal and spatio-temporal point processes. This paper extends the concept to include discrete time when considering its application to long-term patterns of terrorist attacks in Indonesia. The dataset examined here consists of the number of terrorist events per day in Indonesia from 1970 through 2007. The data for 1993 are known to be missing and suspected to be missing for 1998. In order to fully understand the patterns of terrorism over the given time period, the missing data are imputed via Bayesian methods. The computational burden of the model is mitigated by the use of a parallel algorithm for evaluating the likelihood which leads to a substantial gain in computational speed and demonstrates the viability of parallel computation in a broad range of MCMC implementations. Results show that in the immediate aftermath of a terrorist attack, the risk can be increased by over 3400 times the pre-attack level and that after 14 days there is a 90% reduction in the risk of an attack. Policy implications and directions for future research are also discussed.

References

Hamilton, L.C., and Hamilton, J.D. (1983). Dynamics of Terrorism. International Studies Quarterly 27 39-54.

Hawkes, A.G. (1971). Point Spectra of Some Mutually Exciting Point Processes. Journal of the Royal Statistical Society. Series B (Methodological) 33(3) 438-443.

Holden, R.T. (1987). Time Series Analysis of a Contagious Process. Journal of the American Statistical Association 82(400) 1019-1028.

Gentry White *is currently a Research Fellow at the ARC Centre of Excellence in Policing and Security (CEPS) in the Institute for Social Science Research (ISSR) located at the University of Queensland. He received his Ph.D is Statistics from the University of Missouri in 2006 and was a post-doctoral fellow at North Carolina State University and a post-doctoral associate at the Statistical and Mathematical Sciences Institute (SAMSI) from 2006-2009.*

IMPROVED EFFICIENCY IN MULTI-PHASE CASE-CONTROL STUDIES

Alastair Scott[1], Chris J. Wild [2]

[1]Department of Statistics, University of Auckland
38 Princes St, Auckland, New Zealand
aj.scott@auckland.ac.nz

[2]Department of Statistics, University of Auckland
38 Princes St, Auckland, New Zealand
c.wild@auckland.ac.nz

In a recent paper (Lee, Scott & Wild, Biometrika, 2010), we developed efficient methods for fitting regression models to multi-phase case-control data in the special case where all covariates, apart from those measured in the final phase, are categorical. The method can be implemented by taking a linear combination of the estimating equations obtained by applying conditional maximum likelihood (see Breslow & Cain, Biometrika, 1988, 11-20) to the completely-observed units and equations similar to those used in calibrating (see Sarndal, Survey Methodology,2007, 99-119) the sampling weights to quantities known for the partially-observed units. We can extend this procedure to get very good procedures for situations where implementable optimal solutions are not available, for example when continuous covariates are measured in earlier phases or when we have appreciable non-response.

References:
Breslow, N.E. and Cain, K.C. (1988). Logistic regression for two-stage case-control data. Biometrika, 75, 11--20.
Lee, A.J., Scott, A.J. and Wild, C.J. (2010). Efficient estimation in multi-phase case-control studies Biometrika, 97, pp. 361–374, doi: 10.1093/biomet/asq009.
Särndal, C. (2007). The calibration approach in survey theory and practice. Survey Methodology, 33, 99–119.

Professor Chris Wild *- Professor of Statistics at the University of Auckland, New Zealand and recognised by Fellowships of the American Statistical Association and the Royal Society of New Zealand, Chris Wild is a member of a rare crossover species. He publishes extensively in statistical methodology, particularly on response-selective and missing data problems, but also works substantively in statistics education. He co-wrote the Wiley books Nonlinear Regression (1989) and Chance Encounters (2000) with George Seber. His best known statistics education paper is Statistical Thinking in Empirical Enquiry with Maxine Pfannkuch (1999, International Statistical Review). Chris' interests in statistics education include curricular revolution at school levels, growing university statistics programmes, and improving the penetration, quality and practical impact of statistics education at all levels. Chris has been a Council member of the International Statistical Institute, President of the International Association for Statistics Education and an Associate Editor of the International Statistical Review, Biometrics, the Statistics Education Research Journal, and ANZJS. He was Head of Auckland's Department of Statistics 2003-2007 and co-led the University of Auckland's first-year statistics teaching team to a national teaching award in 2003. His keynote addresses include the Royal Statistical Society, the Statistical Society of Canada, and ICOTS.*

INVESTIGATING CAUSAL MECHANISMS OF DISEASE

Elizabeth J. Williamson[1,2], *Elya Moore*[3], *Craig A. Olsson*[3,4]

[1] *Centre for Molecular, Environmental, Melbourne University*
Genetic & Analytic Epidemiology, Melbourne School of Population Health, Parkville, Australia
ewi@unimelb.edu.au

[2] *Department of Epidemiology and Preventive Medicine, Monash University*
The Alfred Centre, 99 Commercial Road, Melbourne, Victoria

[3] *Murdoch Childrens Research Institute*
Royal Children's Hospital, Flemington Road, Melbourne, Victoria
elya.moore@unimelb.edu.au
craig.olsson@rch.org.au

[4] *Psychological Sciences & Paediatrics, University of Melbourne*
Parkville, Melbourne, Victoria

Rothman, in a series of seminal papers in the 1970's (e.g. Rothman, 1976), introduced the concept of disease occurring through a number of sufficient causes, each of which is itself made up of a number of component causes. He used this framework to demonstrate that two risk factors can be said to interact causally on disease occurrence if departure from an additive risk model is observed. This idea has not been widely accepted or implemented in the statistical community in which the great majority of disease modelling is performed on the log-odds scale. The popularity of logistic regression is due to several reasons. Unlike absolute risks, odds ratios have been found to be stable across populations, confounders and non-binary exposure variables can be easily accommodated within a logistic regression framework, and logistic regression can be employed in case-control studies, where no estimate of baseline risk is generally available.

We review the sufficient cause framework of disease causation and discuss its implications for modelling disease. We consider four measures of additive interaction: the Excess Risk due to Additive Interaction (ERI), the Relative Excess Risk due to additive Interaction (RERI), the Attributable Proportion (AP), and the Synergy Index (SI). We show how these can be estimated from regression models on either the log-odds or risk scale. We discuss how confounding and non-categorical exposure variables fit into the framework of sufficient causes.

Theoretical work in this area has concentrated on the case-control design. We therefore consider the issue of obtaining valid statistical inferences for measures of additive interaction in cohort studies and present data from simulation studies concerning the performance of various methods of confidence interval estimation (including the MOVER method (Zou, 2008), the delta method (Hosmer & Lemeshow, 1992) and various bootstrap methods).

References

Hosmer DW & Lemeshow S (1992). Confidence interval estimation of interaction. Epidemiology, 3, 452 – 456.
Rothman KJ (1976). Causes. American Journal of Epidemiology, 104, 587 – 592.
Zou GY (2008). On the estimation of additive interaction by use of the four-by-two table and beyond. American Journal of Epidemiology, 168, 212 – 224.

Elizabeth Williamson *is a Research Fellow working jointly between the Department of Epidemiology and Preventive Medicine at Monash University and at the Centre for MEGA Epidemiology at the Melbourne School of Population Health (MSPH). After completing her PhD in 2007 at the London School of Hygiene & Tropical Medecine she moved to Australia where she has been working in the area of public health and epidemiology. Her main research interests involve the estimation of causal effects from observational data, specifically the estimation of additive interactions within Rothman's sufficient cause framework, propensity score methods, and the use of DAGs in observational studies.*

KINK ESTIMATION IN STOCHASTIC REGRESSION WITH DEPENDENT ERRORS AND PREDICTORS

Justin Wishart[1], *Rafa Kulik*[2]

[1] *University of Sydney, School of Mathematics and Statistics, F07 NSW 2006*
justin.wishart@sydney.edu.au

[2] *University of Ottawa*
rkulik@uottawa.ca

We study the non-parametric estimation of the location of jump points in the first derivative (referred to as kinks) of a regression function μ in a random design model with long-range dependent design points and i.i.d errors. The method is based on the so called zero-crossing technique and makes use of a high-order kernel smoothing approach. The rate of convergence of the estimator is contingent on the level of dependence and the smoothness of the regression function μ.

References

Goldenshluger, A. and Tsybakov, A. and Zeevi, A. (2006). Optimal change-point estimation from indirect observations Ann. Statist., 34 350-372.

Wishart, J. (2009). Kink estimation with correlated noise.. Journal of Korean Statistical Society, 38, 131-143.

Wishart, J. and Kulik, R (2010). Kink estimation in stochastic regression with dependent errors and predictors. Electron. J. Stat. (submitted), URL: http://arxiv.org/abs/1003.1535.

Justin Wishart *is currently a PhD student and Postgraduate Teaching Fellow at The University of Sydney. His area of interest is estimation of non-parametric regression functions and changepoint estimation in both regular design and random design models with possible Long-Range Dependent structures. The preferred estimation methods are using a Kernel smoothing or a Wavelet based approach.*

FITTING HAZARD MODELS BY MAXIMIZING THE LIKELIHOOD UNDER DEPENDENT CENSORING

Jun Ma[1], Kenny Xu[2], Tania Prvan[3]

[1] Department of Statistics, Macquarie University, Australia
jun.ma@mq.edu.au

[2] Department of Statistics, Macquarie University, Australia
jing.xu@mq.edu.au

[3] Department of Statistics, Macquarie University, Australia
tprvan@efs.mq.edu.au

This presentation introduces a novel estimation method for hazard models when censoring time is dependent on the failure time. This dependence is captured by a copula (W.Frees and Valdez (1998)). We use the method of maximum likelihood or maximum penalized likelihood to estimate the hazard or survival functions. In the literature, methods exist for hazard function estimation under dependent censoring, but they maximize the partial likelihood function with an assumed copula (Huang and Zhang (2008)).

Our method is to maximize the full likelihood function which incorporates an assumed copula to model the dependence between censoring and failure times.

References

Edwards W.Frees and Emiliano A. Valdez (1998). Understanding relationships using copulas North American Actuarial Journal, 2, 1-25.

Xuelin Huang and Nan Zhang (2008). Regression Survival Analysis with an Assumed Copula for Dependent Censoring: A Sensitivity Analysis Approach Biometrics, 64, 1090-1099.

Kenny Xu *completed the Bachelor degree of Actuarial Studies and Statistics in 2007. He undertook the honours program in Statistics and was awarded the first class honours in 2008. Kenny continued his PhD study in 2009, in the Department of Statistics at Macquarie University. Kenny's current research interest is hazard function estimation when the censoring time is dependent on the failure time.*

REUSING PRIOR CONTROLS IN NESTED CASE-CONTROL STUDIES: IS IT FEASIBLE?

Qian Yang[1], *Marie Reilly*[2], *Agus Salim*[3]

[1] *Department of Epidemiology and Public Health*
National University of Singapore, Singapore
Yong Loo Lin School Of Medicine, National University of Singapore, MD3, 16 Medical Drive
Singapore 117597
yangqian@nus.edu.sg

[2] *Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Sweden*
Karolinska Institutet, PO Box 281, SE-171 77 Stockholm, Sweden
Marie.Reilly@ki.se

[3] *Department of Epidemiology and Public Health*
National University of Singapore, Singapore
Yong Loo Lin School Of Medicine, National University of Singapore, MD3, 16 Medical Drive
Singapore 117597
ephaguss@nus.edu.sg

Motivation:
Many epidemiological studies use case-control designs to reduce cost and preserve study power. Weighted likelihood methods are available for reusing nested case-control data to supplement a new study. We investigate how the precision of this method varies with the number of controls per case in both the prior study and the new study. As a special case, we explore the feasibility and efficiency of a new study that gathers no controls, relying instead on prior data.

Methods:
We estimated hazard ratios using weighted log-likelihood with the weight given by the inverse of the probability of inclusion in either study. We simulated data to show the relationship between the number of controls per case in each study and precision of the estimate. Using this relationship, we express the contribution of prior controls to a new study in terms of "effective number of controls". We apply our method to a study of anorexia in the Swedish population.

Results:
For a fixed number of controls in the prior study, the relative reduction in the variance decreases as we increase the number of controls in the new study. For a new study with only cases, the estimated variance reduces dramatically as the number of controls in the prior study increases from 1 to 5. We obtained unbiased estimates on applying our method to a study of anorexia cases with no controls, using prior controls from a schizophrenia study to supplement the data.

Conclusion:
We have demonstrated the feasibility of conducting a new study using only incident cases and prior data. The combined analysis of new and prior data gives unbiased estimates of the hazard ratio, with efficiency depending on study size and number of controls per case. This work has important applications in genetic and molecular epidemiology, to make optimal use of costly exposure measurements.

References
Reilly, M., Torrång, A., Klint, Å. (2005). Re-use of case-control data for analysis of new outcome variables. Statistics in Medicine, 24, 4009-4019.
Salim A, Hultman C, Sparén P, Reilly M. Combining data from 2 nested case-control studies of overlapping cohorts to improve efficiency. Biostatistics. 2009 Jan;10(1):70-9.
Samuelsen, S.O. (1997). A pseudo-likelihood approach to analysis of nested case-control studies. Biometrika, 84, 379-394.

Qian Yang *currently works as a research assistant with the Centre for Molecular Epidemiology, Department of Epidemiology and Public Health, in National University of Singapore (NUS). She's also a joint-PhD student with NUS and Karolinska Institutet. Her area of research and study interest is in biostatistical methodology for nested case-controls designs with missing data.*

GENERALISED LINEAR MODELLING OF THE ASTHMA HOSPITALISATION RISK AND AIR POLLUTANT CONCENTRATION IN PERTH

Yano, Y.[1], Mueller, U.[2], Hinwood, A.[3]

[1] Edith Cowan University
100 Joondalup Drive, Joondalup, WA
yyano@ecu.edu.au

[2] Edith Cowan University
100 Joondalup Drive, Joondalup, WA
umueller@ecu.edu.au

[3] Edith Cowan University
100 Joondalup Drive, Joondalup, WA
a.hinwood@ecu.edu.au

Asthma is a respiratory disease which carries a risk of hospitalisation in severe cases. In this study the relationship between air pollutant concentration and asthma hospitalisation risk across Perth's metropolitan area was investigated. Since the number of monitoring stations in the Perth metropolitan area is small, air pollutant concentration estimates were obtained through a combination of a Gaussian Plume Model (GPM) applied to pollutant concentrations at monitoring stations and Lognormal Kriging (LK) of emission inventory data (NO, CO and PM10) for 2006.

The asthma hospitalisation risk was estimated from the monthly recordings of the asthma admissions per postcode, using a Bayesian hierarchical model. The results showed that the asthma hospitalisation risk varies for each postcode with relatively high risk observed in the outskirts of the central metropolitan area. The variable risk modelling was performed in the R statistical computing software using the Poisson-gamma exchangeable prior model

The results of GLM analysis show that there is a statistically significant relationship between asthma hospitalisation risk and air pollutant concentration. The GLM coefficients suggest that an increase in NO and CO concentration results in an increase in the estimated mean risk, while the PM10 concentration shows the opposite relationship. The influence of air pollution on asthma hospitalisation risk is shown to vary over both time and space.

A comparison of the yearly GLM outputs suggests that the asthma hospitalisation risk has decreased over the study period, although a few acute rises in the risk were observed in some years.

Yuichi Yano *completed the Master of Science degree by research and currently works as a Graduate Consultant Statistician at Data Analysis Australia. Yuichi has been involved in various types of projects at Data Analysis Australia since the commencement of his career, and gained experience and further development in data analysis and manipulation skills. His area of interest is in the spatial statistics and Mathematical and Statistical modelling*

ROBUST KALMAN FILTERING WITH MULTIVARIATE GENERALIZED LAPLACE MEASUREMENT NOISE

Pairoj Khawsithiwong[1], Nihal Yatawara[2]

[1]Department of Statistics, Faculty of Science, Silpakorn University
Nakorn Phatom, 73000, Thailand
pairoj@su.ac.th

[2]Department of Mathematics and Statistics, Curtin University of Technology
Perth, Western Australia, 6845, Australia
N.Yatawara@curtin.edu.au

It is well known that the standard Kalman filter fails to provide optimal state estimates in the sense of minimum mean of squared state error, when measurement outliers occur in a linear stochastic system. This is primarily due to the usual Gaussianity assumptions made on the measurement noise term. In this paper, robust Kalman filters are derived using generalized Laplace measurement noise with single and multi scale factors to replace Gaussian assumptions. The performance of the proposed robust filters is compared to the standard Kalman filter through Monte-Carlo simulations.

References

Niehsen, W. (2002). Robust Kalman Filtering with generalized Gaussian measurement noise. IEEE Transactions on Aerospace and Electronic System, 38, 1409-1412.

Pena, D., & Guttman, I. (1988). Bayesian approach to robustifying the Kalman Filter. In Bayesian Analysis of Time Series and Dynamic Models, ed. J.C. Spall, pp. 227-253. New York: Marcel Dekker.

Sorenson, H.W., & Alspach, D.L. (1971). Recursive Bayesian estimation using Gaussian sums. Automatica, 7, 465-479.

Ting, J., Theodorou, E., & Schaal, S. (2007). Learning an outlier-robust Kalman Filter. CLMC Technical Report, University of Southern California, USA.

Yang, Y., & Cui, X. (2008). Adaptively robust Filter with multi adaptive factor. Survey Review, 40, 260-270.

Nihal Yatawara *is currently a lecturer and a researcher at Curtin University of Technology in Perth. His areas of interest are Applications of Time series methods for Quality and process control and Finance. This research is critical to a vast array of practitioners who deal with correlated data where Gaussian assumptions do not hold. Nihal has been involved extensively in consultation, conducting training workshops across several areas of applications of time series analysis for well over 20 years nationally and internationally.*

REDISCOVERING RESULTS: FIRST PASSAGE TIMES FOR SHOT NOISE AND DAMS

Geoffrey Yeo

School of Mathematics and Statistics, University of Western Australia
Nedlands WA 6009
yeo@maths.uwa.edu.au

We outline a first passage time result in dam theory and shot noise. This has appeared in various forms in the same and different journals and been 'rediscovered' several times, with some authors apparently being unaware of its earlier results. As increasing numbers of publications increase the chances of missing earlier work, search engines should decrease it. We explore some of the history of the problem in this context. Further references to those below are in my 2010 paper in The Mathematical Scientist.

References

Kella, O. & Stadje, W. (2001) On hitting times for compound Poisson dams with exponential release rate. J. Appl. Probab., 38, 761-786.
Masoliver, J. (1987) First-passage times for non-Markovian processes: Shot noise. Phys. Rev. A 35, 3918-3928.
Yeo, G. F. (1974) A finite dam with exponential release. J. Appl. Probab., 11, 122-133.

Geoffrey Yeo *is a visiting fellow at UWA, having retired from Murdoch University in 2003. Interests include modelling and analysis of ion channels and storage theory, the secretary problem and characterization problems. My homepage is at www.maths.uwa.edu.au/~yeo.*

ANALYZING CENSORED DATA

Hwan-Jin Yoon

Statistical Consulting Unit, Australian National University
27 John Dedman Building 27, Australian National University, ACT, Canberra 0200, Australia
hwan-jin.yoon@anu.edu.au

The data in designed experiments often contain a large number of zeros. It therefore is clear that the data are not normal. In such cases, this might lead to problems with the application of ANOVA.

If the observed zeros can be regarded as a threshold rather than a true value, censored models might be used. The simplest censored model is the so-called Tobit model introduced by Tobit (1958) as a limited dependent variable model. Guillet et al. (2001) state that the Tobit model is a generalization of ANOVA. But we don't know whether the result from the Tobit model is reliable.

In this paper, we compare the ANOVA with Tobit model using various datasets from agriculture, environment etc. and provide some practical information on how these methods can be assessed.

References

Long J.S. (1997). Regression Models for Categorical and Limited Dependent Variables. SAGE Publications.
Allcroft D.J. and Glasbey C.A. (2003) Analysis of crop lodging using a latent variable model. Journal of Agricultural Science. 140, 383-393.

Jin Yoon *currently works as a statistical consultant in Statistical consulting Unit at ANU, in Canberra.*

AN ALTERNATIVE METHOD FOR FITTING A ZERO INFLATED NEGATIVE BINOMIAL DISTRIBUTION

Z. H. Zamzuri

Department of Statistics
Macquarie University NSW 2109
zamzuri@science.mq.edu.au

There are many techniques proposed in the literature for the fitting of a negative binomial distribution, with their main focus being on estimating of the dispersion parameter. In fitting a zero inflated negative binomial distribution, an additional parameter; π must be estimated, the proportion of extra zeros. In this presentation, we describe a method for fitting zero inflated negative binomial distribution based on conditional maximum likelihood and a simple grid search. We compare results of the goodness of fit test of this method with that of the 'zeroinfl' program in R that is based on maximum likelihood. Using simulation, it is shown that the alternative method offers a better fit when the data is highly dispersed.

References

K. Anraku and T. Yanagimoto (1990). Estimation for the negative binomial distribution based on the conditional likelihood.
Communications in Statistics – Simulation and Computation, 19:3, 771-786.
A. Zeileis, C. Kleiber and S. Jackman (2008). Regression model for Count Data in R.
Journal of Statistical Software 27:8.

Zamira Zamzuri *currently is a Ph. D student in Statistics at Macquarie University under Professor Graham Wood's supervision. Her area of interest is the application of statistical modelling in traffic accidents. Zamira started her study in Macquarie University last year.*

HOW MUCH RISK ASSOCIATED WITH A BIOMARKER CAN BE ACCOUNTED FOR BY OTHER PROGNOSTIC VARIABLES IN TIME-TO-EVENT OUTCOMES?

[1] Zannino D, [2] Byth K, [3] Ting R, [4] Keech A, [5] Gebski V

[1] NHMRC Clinical Trials Centre
Locked Bag 77, Camperdown 1450, NSW
Diana.Zannino@ctc.usyd.edu.au

[2] NHMRC Clinical Trials Centre
Locked Bag 77, Camperdown 1450, NSW
Karen.Byth@ctc.usyd.edu.au

[3] NHMRC Clinical Trials Centre
Locked Bag 77, Camperdown 1450, NSW
Rudee.Ting@ctc.usyd.edu.au

[4] NHMRC Clinical Trials Centre
Locked Bag 77, Camperdown 1450, NSW
Tony@ctc.usyd.edu.au

[5] NHMRC Clinical Trials Centre
Locked Bag 77, Camperdown 1450, NSW
Val@ctc.usyd.edu.au

With the proliferation of biomarkers identified in medical research, the importance of individual (or a class of) biomarkers over other prognostic factors is of major interest. In a proportional hazards (PH) regression model, the joint association between variables in the model and the dichotomous event of interest, can be summarised in terms of global performance by the c-index (Harrell-Lee). This index is analogous to the area under the Receiver Operating Curve but allows for censored data. This idea is used to estimate the proportion of a biomarker effect which is accounted for by a set of other prognostic variables by examining changes in the c-index under different PH models.

The identification of key prognostic variables may be determined through an exhaustive search subset regression. Provided the different models produced by the exhaustive search are nested, changes in the c-index associated with the successive addition of each variable can be obtained. The percentage change from the c-index, containing just the biomarker of interest is calculated. This approach was applied to a large RCT in diabetes to determine the proportion of eGFR (renal function) prognostic information which is explained by other key variables (HDL cholesterol, HbA1c, etc) on cardiovascular risk.

The use of the c-index in conjunction with an exhaustive search subset regression provides an easily implemented and intuitively appealing method of quantifying the proportion of a 'biomarker's effect' explained by other prognostic variables. This can then be used to help assess the clinical value of different biomarkers, eliminating those whose prognostic effect is explained by more conventional and readily available clinical tests.

References

Harrell, F.E., Califf, R.M., Pryor, D.B., Lee, K.L., and Rosati, R.A. (1984). 'Evaluating the Yield of Medical tests", Journal of the American Medical Association, 247(18), pp.2543-2546
Harrell, F.E, Lee, K.L, Califf, R.M., Pryor, D.B., and Rosati, R.A. (1984), 'Regression Modelling Strategies for Improved Prognostic Prediction", Statistics in Medicine, 3(2), pp.143-152.

Diana Zannino *currently works as a Biostatistician with the NHMRC Clinical Trials Centre, University of Sydney, in NSW. She works on clinical trials in cardiovascular and oncology medicine and assists teaching in the Biostatistics Collaboration of Australia (BCA) program and MPH in the School of Public Health at the University of Sydney.*

VARIANCE/BANDWIDTH SELECTION IN NORMAL MIXTURE DENSITY ESTIMATION

M. Stewart[1], J. Zhu[2]

[1] Sydney University, School of Mathematics and Statistics F07 University of Sydney NSW 2006 Australia,
michaels@maths.usyd.edu.au

[2] Sydney Universiry, School of Mathematics and Statistics F07 University of Sydney NSW 2006 Australia,
jennyz@maths.usyd.edu.au

Fitting an arbitrary normal location mixture with an unknown, fixed variance can be viewed as a generalistion of the well-known normal kernel density estimation procedure. The unknown common variance plays the role of the bandwidth, and choosing it is a difficult problem in both contexts. In the mixture context however, it is a parameter that can in principle be estimated, an option which is not generally available in the kernel-estimation context. We shall discuss various theoretical and computational aspects of this variance/bandwith selection procedure, including properties of the nonparametric maximum likelihood estimator of the mixing distribution (Lindsay, 1983) and some simulation studies.

References

Lindsay, B.G.
(1993). The geometry of mixture likelihoods: A general theory. The Annals of Statistics, 11, 86-94.

Jennifer Zhu *is currently a PhD student at the University of Sydney Department of Statistics. Her main interests are mixture models and computational statistics.*

USING RECORD LINKAGE TO SURVEY DATA TO CORRECT FOR UNDER-IDENTIFICATION OF ABORIGINAL BIRTHS ON THE WA BIRTH REGISTER

Zubrick S R[1], Mitrou F[2], Lawrence D[3], Christensen D[4], Hancock K[5], Hafekost K[6]

[1] Centre for Developmental Health
Curtin University of Technology and Telethon Institute for Child Health Research
PO Box 855, West Perth, WA. 6872
S.Zubrick@curtin.edu.au

[2] Telethon Institute for Child Health Research
GPO Box K881, Perth, WA. 6842
francism@ichr.uwa.edu.au

[3] Centre for Developmental Health
Curtin University of Technology and Telethon Institute for Child Health Research
GPO Box K881, Perth, WA. 6842
D.Lawrence@curtin.edu.au

[4] Telethon Institute for Child Health Research, GPO Box K881, Perth, WA. 6842
dchristensen@ichr.uwa.edu.au

[5] Telethon Institute for Child Health Research, GPO Box K881, Perth, WA. 6842
khancock@ichr.uwa.edu.au

[6] Telethon Institute for Child Health Research, GPO Box K881, Perth, WA. 6842
khafekost@ichr.uwa.edu.au

Administrative registers are routinely used to produce time series of vital statistics for Aboriginal and non-Aboriginal populations. With the Australian Government's commitment to Closing the Gap, the quality of ascertainment of Aboriginal status on registers has become a topical issue. We used data from the Western Australian Aboriginal Child Health Survey (WAACHS, Zubrick et al., 2004), to evaluate the identification of Aboriginal births on the WA birth register. The WAACHS was a population-based probability sample of 5,300 Aboriginal children aged under 18 years, and their families, drawn from across the state, representing a sampling fraction of about 1 in 6. Where consent was given (around 96%), we linked each survey child to the WA birth register. We found that around 20% of Aboriginal children in the survey were not identified as Aboriginal on the birth register. Of these around two-thirds had a non-Aboriginal mother and an Aboriginal father, and the remainder the mother identified as Aboriginal in the survey but was not recorded as Aboriginal on the birth register. Characteristics of births identified as Aboriginal in both sources were different from those identified as Aboriginal only in the survey, with a higher proportion of births identified as Aboriginal in both sources being of low birth weight or premature. Using knowledge of the differences in characteristics between these two sets of births we developed a correction to time series of proportion of premature and low birth weight Aboriginal babies derived from the WA birth register. Being born with low birth weight or premature are important constraints on human capability development. Record linkage of administrative data to survey data sets may provide a methodological tool for assessing the quality of administrative data and improving the accuracy of series based on administrative data.

References

Zubrick S.R., Lawrence D., Silburn S.R., Blair E., Milroy H., Wilkes T., Eades S., D'Antoine H., Read A., Ischiguchi P., Doyle S. (2004). The Western Australian Aboriginal Child Health Survey: The health of Aboriginal children and young people. Perth: Telethon Institute for Child Health Research.

Steve Zubrick *a Professor at Curtin and Head of Population Sciences at the Telethon Institute for Child Health Research, completed his doctoral and postdoctoral work in psychology at The University of Michigan and worked in mental health settings for many years before starting at the Institute in 1991. His research interests include the social determinants of health and mental health in children, systematic studies of youth suicide, and large scale psychosocial survey work in non-Indigenous and Indigenous populations. Professor Zubrick is considered a leading Australian authority in the epidemiology of child and adolescent mental health and in public health approaches to promotion and prevention of mental health. He chairs the Consortium Advisory Group of the Longitudinal Study of Australian Children and featured in the ABC TV's 'Life at' documentary series.*

STATISTICAL MODELLING AND ANALYSIS OF ION CHANNEL DATA

Ibrahim Almanjahie[1], *Robin Milne*[2], *Nazim Khan*[3]

[1] *School of Mathematics and Statistics, University of Western Australia*
Crawley 6009
ibrahim@maths.uwa.edu.au

[2] *School of Mathematics and Statistics, University of Western Australia*
Crawley 6009
milne@maths.uwa.edu.au

[3] *School of Mathematics and Statistics, University of Western Australia*
Crawley 6009
nazim@maths.uwa.edu.au

Ion channels are specialised protein molecules which selectively control the movement of ions and molecules across biological membranes, thereby regulating many aspects of cell function. Passage of the ions and molecules occurs through aqueous pores gated usually in an all-or-none fashion by specific stimuli, including mechanical stress. The refined patch clamp technique (Hamill et al., 1981) allows channel ionic currents to be recorded.

One type of channel is the mechanosensitive large conductance channel (MscL) in the bacterium E. Coli. Determination of the structure of MscL and developments in cloning capability (Sukharev et al., 1994) have enabled unprecedented study of MscL functional behaviour. In turn, the extensive high quality patch clamp data allows model building and testing of basic mechanisms of channel kinetics.

Analysis of patch clamp data from MscL in E. Coli has been limited due to a number of challenges, such as the multi-level nature of the current and suitable recording bandwidth. In my research, I will focus on analysis of higher bandwidth (100 kHz) patch clamp data from MscL using hidden Markov models (HMMs) incorporating correlated noise. The aim is to determine the number of conductance levels of the channel, together with mean current, mean dwell time and equilibrium probability of occupancy for each level.

References

Hamill, O. P., Marty, A., Neher, E., Sakmann, B. and Sigworth, F. J. (1981). Improved patch-clamp techniques for high-resolution current recording from cells and cell-free membrane patches. Pflugers Archiv - European J. Physiology, 391, 85-100.
Khan, R. N., Martinac, B., Madsen, B. W., Milne, R. K., Yeo, G. F. and Edeson, R. O.(2005). Hidden Markov analysis of mechanosensitive ion channel gating. Mathematical Biosciences, 193, 139-158.
Sukharev, S. I., Blount, P., Martinac, B., Blattner, F. R. and Kung, C. (1994). A large-conductance mechanosensitive channel in E. coli encoded by mscL alone. Nature, 368, 265-268.

Ibrahim Almanjahie *received the B.Sc. degree in Mathematics from the King Khalid University, Saudi Arabia in 2002. He received the M.Sc. degree in mathematical and statistical science from the University of Western Australia, Australia in 2009. Ibrahim Almanjahie is currently a PhD candidate at the School of Mathematics and Statistics, the University of Western Australia. His major research interests are: applied statistics, stochastic analysis and modelling and ion channel modelling.*

EFFICIENCY OF USING SPATIAL ANALYSIS IN MULTI-ENVIRONMENTAL TRIALS

Ibrahim Alsayed

Department of Biometry and Experimental Design, Humboldt-Universität zu Berlin
Faculty of Agriculture and Horticulture, 101115 Berlin, Invalidenstr.42, Germany
syriadeutsch@hotmail.com

Randomised complete block design (RCB) and incomplete block design (IB) are techniques used to control the spatial variation in field trials. Spatial dependency between the plots is not considered. Spatial models take into account this dependency. In variety trials with large blocks, spatial heterogeneity within the field may reduce the efficiency of the genotype selection. Grain yield data from a series of trials performed within a breeding program were analyzed using 24 spatial models to compare the conventional methods with geostatistical methods. Akaike's Information Criterion (AIC) and the Likelihood-ratio test (LRT) were used to assess the best spatial model. It was found that more than 80% of the data sets showed significant model improvement with spatial analysis. The relative efficiency improved up to 250% for the best spatial model. There was noticeable change in ranking of genotypes.

Ibrahim Alsayed *is currently working as a PhD Student at the division of Biometry and Experimentation in the faculty of Agriculture and Horticulture, Humboldt-Universität, Berlin. My area of interest is spatial and temporal analysis of field trials. I worked for the General Commission of Scientific Agricultural Research (GCSAR) in Syria from 2001-2005.*

MODELLING PRESENCE OF RARE SPECIES IN AUSTRALIA'S EXCLUSIVE ECONOMIC ZONE

David Clifford[1], Keith R. Hayes[1], Chris Moeseneder[2], Mark Palmer[1] and Tom Taranto[2]

[1]CSIRO Mathematics, Informatics and Statistics
Longpocket Laboratories, Indooroopilly, QLD 4068
David.Clifford@csiro.au

[2]CSIRO Marine and Atmospheric Research
PO Box 120, Cleveland, QLD 4163

In 2009 CSIRO and the Department of Environmental, Water Heritage and the Arts, embarked on a landmark study to capture data on the environmental assets and threats present in Australia's Exclusive Economic Zone (EEZ). Data was collated from many federal and state agencies on issues as broad as recreational and commercial fishing, oil exploration, key ecological features (KEFs) and marine biodiversity. A significant part of the project involved mapping where threatened, endangered and protected species are likely to be found within the EEZ. Presence-only data, often recorded in a non-systematic way, are available for many species. Systematically collected presence/absence data are also available from independent sources for some species.

In this presentation we outline our work on mapping the assets and threats. The ultimate goal is to predict the effect of anthropogenic pressures on KEFs as these are considered to have a disproportionately large effect on important ecosystem processes. We highlight progress made using presence-only data to model the distribution of species. Logistic regression (Ward et al, 2009), rule ensembles (Friedman and Popescu, 2008) and maxent (Phillips et al 2004) are three common approaches used for modelling data of this type. The results of their application to this problem will be presented and problems and pitfalls in dealing the presence only data will be outlined.

References

Friedman, J. H. and Popescu, B. E. (2008). Predictive learning via rule ensembles. Annals of Applied Statistics, 2(3):916–954.

Phillips, S. J., Dudik, M., and Schapire, R. E. (2004). A maximum entropy approach to species distribution modeling. In Proceedings of the Twenty-First International Conference on Machine Learning, pages 655–662.

Ward, G., Hastie, T., Barry, S., Elith, J., and Leathwick, J. (2009). Presence-only data and the EM algorithm. Biometrics, 65:554–563.

David Clifford *is a research scientist at CSIRO Mathematics, Informatics and Statistics, and is originally from Cork, Ireland. David has a PhD in statistics from the University of Chicago for his research on the nature of spatial variation in crop yields under the guidance of Prof Peter McCullagh. David's research and work in statistics has been driven by applied and computational problems.*

STRUCTURAL EQUATION MODELLING IN BUSINESS PROBLEMS

Emel Şıklar[1], Duygu Çoşkun[2]

[1] Anadolu University
Faculty of Economics and Administrative Sciences, Anadolu University, 26470 Eskişehir, Turkey
esiklar@anadolu.edu.tr

[2] Anadolu University
Faculty of Economics and Administrative Sciences, Anadolu University, 26470 Eskişehir, Turkey
dcoskun@anadolu.edu.tr

Structural Equation Modelling (SEM) is a widely used data analysis technique in social science investigations. In this technique the linear relationships, either directly observable or unobservable, between multiple dependent and independent variables are analyzed simultaneously. Various theoretical models can be tested in SEM that hypothesize how sets of variables define constructs and how these constructs are related to each other. The goal of SEM is to determine the extent to which the theoretical model is supported by sample data.

In this study, after a brief explanation of SEM, an application of this technique is given which tries to show the relations between structures of interest.

References

Jöreskog, K.G., and Sörbom, D. (1993). Lisrel:8 Structural Equation Modeling with the SIMPLIS Command Language, USA.: Scientific Software International,Inc.
Schumacker, R.E. and Lomax R.G. (2004). A Beginner's Guide to Structural Equation Modelig,London: Lawrence Erlbaum Associates, Inc.
Raykov T. and Marcoulides, G.A. (2006). A First Course in Structural Equation Modeling,
London: Lawrence Erlbaum Associates, Inc.

Duygu Çoşkun *currently works as a research assistant in Anadolu University Department of Quantitative Sciences in Business Administration. Duygu got her undergraduate degree from the department of Statistics in 2004. She has been doing her PhD. since 2007. Her PhD. thesis is about Structural Equation Modelling. She attended some developmental programs about statistics (Lisrel 8. and etc.)*

MULTIVARIATE DATA VISUALIZATION: CONSTELLATION GRAPH REVISITED WITH EXAMPLES

Siva Ganesh[1] and Selvanayagam Ganesalingam[2]

[1]Massey University
Inst. of Fundamental Sciences, Private Bag 11222, Palmerston North 4442, New Zealand
s.ganesh@massey.ac.nz

[2]Massey University
Inst. of Fundamental Sciences, Private Bag 11222, Palmerston North 4442, New Zealand
s.ganesalingam@massey.ac.nz

Visualization of multivariate data is an important aspect of exploring similarity among observations. Traditionally, multidimensional scaling (MDS) based on proximity measures such as Euclidean distance provided a useful tool for visualization of high-dimensional data in lower dimensions. Techniques such as principal component analysis (PCA) can also be used for displaying high-dimensional data in a low-dimensional space. MDS preserves the pair-wise distance between the data points while representing the embedding in a low-dimensional coordinate system. PCA is a dimensionality reduction method that preserves the overall variation in the data when representing it in a low dimensional space. In both cases, there may be (substantial) loss of information when projecting data onto a 2-dimensional space.

Constellation graph (CG) can be regarded as a process that provides an efficient way of displaying multivariate data in a 2-dimensional space without loss of information. As opposed to data being displayed on perpendicularly intersecting planes in the MDS and PCA approaches, in the CG approach each data point is positioned in the upper half of a unit circle using the polar coordinate system. In this paper, we revisit the basic concepts of CG and demonstrate its advantages via some real world examples involving regression and classification (with class-imbalance) problems.

References

Wakimoyo, K., and Taguri, M. (1997). Constellation graphical method for representing multi-dimensional data, Ann. Inst. Statist.Math., 30, 97-104.

Siva Ganesh *is a Senior Academic in Statistics at the Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand. He has specific expertise in Applied Statistics, Data Mining and Statistical Computing and Consulting. He has been involved in university teaching including running workshops and short courses in New Zealand and internationally, in the areas of multivariate statistics, experimental designs and analysis, biostatistics and data mining. His current research interests include, absolute-value discriminant analysis, classification problems involving class-imbalance, feature selection, clustering and visualization, and predicting RNA-infrastructure via visualizing integrated regulated protein networks.*

QUALITY CONSIDERATIONS IN A PROCESS CONTROL SURVEY – EXPLORATION OF THE GOVERNANCE DIMENSION

Stephen Horn [1], David Lawrence[2]

[1] Dept of Families, Housing, Community Services and Indigenous Affairs
Tuggeranong Office Park, Athllon Drive, Greenway, ACT
stephen.horn@fahcsia.gov.au

[2] Dept of Families, Housing, Community Services and Indigenous Affairs
Tuggeranong Office Park, Athllon Drive, Greenway, ACT
david.lawrence@fahcsia.gov.au

How closely field staff are implicated in survey outcomes is not a usual cause of concern for methodologists. We study the bias from discretionary action of field staff contracted to undertake a process control survey, but otherwise employed by the agency responsible for the process, as a component in the quality assurance surrounding publication of payment accuracy statements based on survey estimates.

The random sample surveys system (RSS) is used across government portfolios to measure accuracy and reliability in the delivery of welfare payments. Eligibility for payment of randomly selected welfare customers is tested, usually by personal interview. Customer service officers, specially selected and trained, undertake these reviews. Inevitably review results reflect both on the design of payments and on the administrative process approving and setting payment. Are controls sufficiently developed to quarantine CSO-agency bias? De-selection is one source of discretionary bias. It is examined using access to administrative history of selected customers. There is prima facie reason to both implicate and discount bias; our results lend weight to the latter.

References

Horn, S. (2008) Quality Assurance Surveys and Program Administration – when accuracy really counts, contributed paper European Conference on Quality in Official Statistics, Rome

Stephen Horn *works as a statistician in the Social Security Policy Branch of the Australian Government Department of Families, Housing, Community Services and Indigenous Affairs. He is interested in the collection design, management, quality assessment and use of longitudinal data within government. He has contributed to forums on nonresponse and adjustmernt of household and customer surveys using experience gained at his current post and when working previously in the methods division of the Australian Bureau of Statistics.*

GOODNESS OF FIT TEST UNDER SRS AND RSS FOR LOGISTIC DISTRIBUTION USING KULLBACK-LEIBLER INFORMATION

S. A. Al-Subh[1], K. Ibrahim[2], A. A. Jemain[3], M. T. Alodat[4],

[1,2,3]School of Mathematical Sciences, Universiti Kebangsaan Malaysia, Selangor, Malaysia
[1]salsubh@yahoo.com, [2]Kamarulz@ukm.my, [3]azizj@ukm.my

[4]Department of Statistics, Yarmouk University, Irbid, Jordan
[4]alodatmts@yahoo.com

Let $X_1, X_2, ..., X_n$ be independent observations on a random variable X. In this paper, our objective is to test the statistical hypothesis that the unknown distribution of X belongs to the logistic distribution. In this study, a goodness of fit test statistics for the logistic distribution based on Kullback-Leibler information is studied. The logistic parameters are estimated using several methods of estimation such as maximum likelihood, order statistics, moments, L-moments and LQ-moments. The critical value based on the statistic which involves the Kullback-Leibler information under the assumption that H_0 is true is computed using Monte Carlo simulations.

The performance of the test under ranked set sampling is investigated. Ten different distributions are considered under the alternative hypothesis. Based on Monte Carlo simulations, for all the distributions considered, it is generally found that the test statistics based on estimators found by moment and LQ-moment methods have the highest power.

Kamarulzaman Ibrahim *is a lecturer in the Statistics Program, School of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia. He is actively involved in teaching statistics courses and research in the school. His research interests include statistical modelling, statistical inference and sampling.*

ROBUST ESTIMATION OF PARAMETERS OF STABLE DISTRIBUTIONS

Eliud Kangogo[1], *Andrzej S. Kozek*[2]

[1] *Macquarie University*
Eliud.Kangogo@mq.edu.au

[2] *Macquarie University*
Andrzej.Kozek@mq.edu.au

Over recent years there has been great interest in using alpha stable distributions for modelling heavy-tailed data in Finance. Despite their attractive properties of stability and the generalized Central Limit Theorem, the lack of a closed form expression for the density function contributes to the difficulty in estimation of their parameters. Although many researchers have addressed this problem, many questions about robustness of estimators remain so far unanswered. Maximum likelihood estimators for example, though efficient, are known to be non-robust, cf. (Huber 1981).

Basu et al. (1998) introduced a new family of density-based divergence measures and showed that for some parametric families the new minimum discrepancy estimators of parameters have nice robustness properties. In the present project we explore robustness properties of similar minimum discrepancy estimators of parameters of stable distribution functions. We derive influence functions for the proposed statistical functionals and the related asymptotic variances of estimators of parameters. We also explore other robustness properties of the minimum discrepancy estimators and illustrate via simulations their performance in finite samples.

References

Basu, A., Harris, I. R., Hjort, N.L., and Jones, M.C. (1998). Robust and efficient estimation by minimising a density power divergence. Biometrika, 85, 549-559.
Huber, P. J. (1981). Robust Statistics. Wiley, New York.

Eliud Kangogo *is currently a PhD student in the Department of Statistics at Macquarie University, Sydney. He holds a Master's degree in Financial Mathematics from the University of Kaiserslautern, Germany and a Bachelor's degree (Hons) from Jomo Kenyatta University of Agriculture and Technology (JKUAT), Kenya. His research interests are in quantitative finance and stochastic modelling. His PhD dissertation concentrates on robust statistical approaches for estimating heavy-tailed distributions useful for modelling returns from financial markets. He also taught at JKUAT in 2007.*

THE IMPACT OF MISSING DATA ON ANALYSES OF TIME-DEPENDENT COVARIATES IN A LONGITUDINAL COHORT: A SIMULATION STUDY

Karahalios E[1], Baglietto L[2], English DR[3], Simpson JA[4]

[1] Cancer Epidemiology Centre, Cancer Council Victoria
Centre for Molecular, Environmental, Genetic and Analytic Epidemiology, University of Melbourne
1 Rathdowne Street, Carlton, Victoria, 3053
Emily.karahalios@cancervic.org.au

[2] Cancer Epidemiology Centre, Cancer Council Victoria, Melbourne, Australia
Centre for Molecular, Environmental, Genetic and Analytic Epidemiology, University of Melbourne
1 Rathdowne Street, Carlton, Victoria, 3053
laura.baglietto@cancervic.org.au

[3] Centre for Molecular, Environmental, Genetic and Analytic Epidemiology, University of Melbourne
Cancer Epidemiology Centre, Cancer Council Victoria, Melbourne, Australia
Level 1 / 723 Swanston St, Carlton, Victoria, 3010
d.english@unimelb.edu.au

[4] Centre for Molecular, Environmental, Genetic and Analytic Epidemiology, University of Melbourne
Cancer Epidemiology Centre, Cancer Council Victoria, Melbourne, Australia
Level 1 / 723 Swanston St, Carlton, Victoria, 3010
julieas@unimelb.edu.au

Missing data are an inevitable source of problems in longitudinal cohort studies, especially cohorts that follow up participants over long periods of time. The most popular method used to handle missing data is complete case analysis, which excludes participants with any missing values from the analysis. An alternative approach is multiple imputation, where each missing value is replaced with multiple plausible values imputed from a statistical model. The multiple completed datasets are then analyzed separately and the resulting estimates combined to obtain a single estimate(Rubin, 1987). Whether it is better to use multiple imputation for handling missing time-dependent covariates or to perform a complete-case analysis cannot be answered with a general rule of thumb. Instead the two methods need to be compared for a variety of scenarios using simulation studies.

We have performed a simulation study to compare multiple imputation and complete-case analysis for dealing with missing data in an analysis of the association between a time-dependent covariate (waist circumference) measured at two time points (baseline and 13 years follow up) and a time-to-event outcome (colorectal cancer) adjusted for confounders. Using data from the Melbourne Collaborative Cohort Study (MCCS) (Giles and English, 2002), we have generated datasets to resemble the distribution of the time-dependent covariates (waist circumference and confounders), their interrelationships and the outcome using the correlation structures of the observed data. Varying proportions of missing data (5%, 15%, 35% - as observed in MCCS, 50%) were imposed on the follow-up measurements of the primary time-dependent covariate (i.e. waist circumference). Three different mechanisms for missingness were chosen: missing completely at random, missing due to confounders measured at baseline (country of birth, age, education level and physical activity), and missing due to the potentially unobserved covariate, waist circumference.

References

Giles, G. G. and English, D. R. (2002). The Melbourne Collaborative Cohort Study. IARC Sci Publ, 15, 69-70.
Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys, New York: John Wiley & Sons.

Emily Karahalios *currently works in the Cancer Epidemiology Centre at the Cancer Council Victoria and is completing her PhD through the Centre for Molecular, Environmental, Genetic and Analytic Epidemiology at the University of Melbourne. Emily's PhD is looking at the effect of drop out in a longitudinal cohort study on the association between change in waist circumference and cancer risk.*

YULE'S CHARACTERISTIC- K

Kumud Gore Kherdekar[1], S.G. Prabhu Ajgaonkar[2]

[1]Govt. college of Arts & Science
Aurangabad,(Maharashtra),India
kumudanil@gmail.com

[2]Dr. B. A.M. University,Aurangabad(Maharashtra), India

Word distribution is a typical type of distribution, wherein the form of the distribution changes with the change in sample size. It belongs to the class of distributions of "multiple happenings". Word distribution is a frequency distribution of frequencies. Hence it becomes difficult to compare works of two different authors or two works of the same author.

Therefore some statistic which characterizes the word distribution and yet is independent of sample size is necessary. Udny Yule(1944) has proposed a statistic named as Yule's Characteristic K for this purpose. This research paper is an attempt to examine if Yule's K is independent of sample size.

For this study we have considered a novel "Tess Of D'Urbervillies" by famous English author Thomas Hardy.

For various samples of different sizes Yule's Characteristic K was obtained and the data was analysed by using Karl Pearson's coefficient of correlation and graphs.

Kumud Gore Kherdekar *is presently working as Associate Professor & Head of department of Statistics, in Govt. College of Arts & Science,Aurangabad,Maharashtra,India. She is also working as a National Cadet Core officer, which is a voluntary organization for youth. Her area of research is Statistical Linguistics and of interest is Applied Statistics. She is actively involved in consultation by MBA Ph.D.,MBBS and MDS students and she also works on Research Projects.*

PROBABILISTIC SENSITIVITY ANALYSIS OF RELATIVE RISK IN HEALTH ECONOMIC MODELLING: WHICH DISTRIBUTION TO CHOOSE?

Hansoo Kim[1], Danny Liew[2], Lyle Gurrin[3]

[1] Department of Medicine The University of Melbourne
29 Regent Street Fitzroy, Melbourne, VIC 3065
h.kim14@pgrad.unimelb.edu.au

[2] Department of Medicine The University of Melbourne
29 Regent Street Fitzroy, Melbourne, VIC 3065
dliew@medstv.unimelb.edu.au

[3] School of Public Health The University of Melbourne
723 Swanston Street Carlton, VIC 3053
lgurrin@unimelb.edu.au

Probabilistic sensitivity analysis (PSA) is important in health economic modelling. PSA is done by assigning distributions to the input parameters and subsequently performing simulations. Each simulation will then represent an economic evaluation. One objection to the PSA technique is that the choice of distributions is subjective and can therefore potentially be manipulated for a favorable outcome.

The triangular distribution is often used, when limited information is available. In this study the triangular, lognormal and empirical distribution were used to represent the distribution of the relative risk (RR).

Parameters for the simulations were estimated using data on responder proportions from a clinical trial. Responder rates in the clinical trial were 492/692=0.71 for treatment A and 449/669=0.67 for treatment B. This resulted in an observed RR of 1.06.

Ten thousand simulations of the triangular distribution, the lognormal distribution and relative risks constructed from two binomial distributions were performed (treatment A was assumed independent from treatment B). All parameters in the simulated distributions were estimated using the same data from the clinical trial. Descriptive statistics and graphical plots were constructed.

The relative risk was overestimated by the triangular distribution (1.57) and the lognormal distribution (1.52). The empirical distribution derived from simulations of two separate binomial distributions was1.06.

This study shows that the triangular distribution is a poor choice for characterizing the uncertainty of RR. The overestimation of the RR can introduce bias for instance if used for responder rates leading to a potentially more favorable outcome. The lognormal distribution appears to be a slightly better approximation, however in this example it overestimates the RR. If the actual number of events and total number of exposed are available the empirical simulation is preferred.

Hansoo Kim *is an accredited statistician with degrees in mathematical and theoretical statistics from the University of Copenhagen, Denmark. He is an experienced clinical statistician and health economist. Currently Hansoo is doing an NHMRC funded project on economic and epidemiological modelling in diabetes at the University of Melbourne.*

A TRANSITION MODEL FOR ORDINAL RESPONSE RHEUMATOID ARTHRITIS DATA

Nursel Koyuncu[1], *Emmanuel Lesaffre*[2]

[1] *Hacettepe University, Faculty of Science, Department of Statistics*
06800, Beytepe, Ankara, TURKEY
nkoyuncu@hacettepe.edu.tr

[2]*Erasmus University Rotterdam, Erasmus Medical Centre*
Department of Biostatistics, 3000, Rotterdam, the Netherlands

The Rotterdam Early Arthritis (REACH) study is a prospective ongoing cohort study set up in 2004. It aims to identify rheumatoid arthritis patients as early as possible. In REACH the patients are diagnosed and treated by rheumatologists. In the first 4 years of the study no guidelines were given and doctors could treat patients using what they thought would be the best method in their own opinion. In this study we need to deal with unprotocolised treatment. The data structure has a repeated measurements flavour. The dose for each patient is increased depending on a value scored by the treating physician. For this study we use a transition model to investigate the treatment effect on patients suffering from rheumatoid arthritis.

References

Ghahroodi, Z.R., Ganjali, M., Berridge, D. (2009) A transition model for ordinal response data with random dropout: An Application to fluvoxamine data, Journal of Biopharmaceutical Statistics, 19, 658-671.

Yu, F., Morgenstern, H., Hurwitz E., Berlin., T.R. (2003) Use of a markov transition model to analyse longitudinal low-back pain data, Statistical Methods in Medical Research, 12, 321-331.

Nursel Koyuncu *currently works as a research assistant at the Statistics Department in Hacettepe University. She also studied for seven months in the REACH study at Erasmus MC medical Center as a researcher. Her area is survey sampling and biostatistics.*

SATISFACTION OF SOUTH AFRICANS REGARDING PROVIDED HEALTH CARE SERVICES

Collen Makomane

Human Science Research Council, Private Bag X41, Pretoria 0001, South Africa
cmakomane@hsrc.ac.za

Patient satisfaction has been the concern in policy on quality in health care for South Africans. The policy was developed to regulate the re-engineering of internal operations in health care services, to improve efficiency, effectiveness, quality and satisfaction. The policy confirmed that patient satisfaction surveys will be a mechanism to measure patient satisfaction. Hence the aim of this study is to track annual changes of patient satisfaction using the annually conducted survey General Household Survey (GHS), to compensate for the shortage of data from other years and to asses the capability of GHS to measure satisfaction of patients. Data will be extracted from GHS (2002-2008) datasets considering only respondents who answered 'yes' on the question about their consultation with a health worker. Means, Standard Deviation, Independent sample T-Test, Simple Regression and the Scheffe Test will be used to answer the study's main questions.

Collen Makomane *is a senior programmer at the Human Science Research Council (HSRC) with more than 6 years experience in statistical data analysis. He holds BSc Honors in statistics from University of Limpopo. Before joining the HSRC in August 2009, he was assistant director of the Health Research unit at the National Department of Health. The Health Research unit was responsible for the coordination of research projects. His areas of research interest are large sample surveys analysis, quantitative data management, data mining, statistical analysis, and monitoring & evaluation.*

BAYESIAN INFERENCE AND PREDICTION IN AN M/G/1 QUEUE WITH OPTIONAL SECOND SERVICE WITH TWO DISCIPLINES

A. R. Mohammadi[1], *M. R. Salehi-Rad*[2]

[1] *Department of statistics, Allameh Tabataba`i University, Tehran, Iran*
r_627@yahoo.com

[2]*Department of statistics, Allameh Tabataba`i University, Tehran, Iran*
moresara20@yahoo.com

This paper describes a Bayesian approach to make inference and prediction for an M/G/1 queueing system with optional second service in which a production item is failed with probability p and it is then repaired with two disciplines I and II. First, a flexible model based on mixtures of Erlang distributions is proposed to approximate the service and reservice time densities. Then, given the sample data, we propose a Bayesian procedure based on a birth-death MCMC method to estimate some performance measures for our system. The approach is illustrated with a real data.

References

Ausin, M.C.,Wiper, M.P., Lillo, R.E., (2004). Bayesian estimation for the M/G/1 queue using a phase type approximation. Journal of Statistical Planning and Inference, 118, 83–101.

Salehi-Rad, M.R., Mengersen, K., (2002). Reservicing some customers in M/G/1 queues, under two disciplines, Advances in Statestics, Combinatorics and Related Areas, World Scientific Publishing, New Jersey, pp. 267-274.

Salehi-Rad, M.R., Mengersen, K., Shahkar, G. H., (2004). Reservicing some customers in M/G/1 queues, under three disciplines, International Journal of mathematics and mathematical sciences, pp. 1715-1723.

Abdolreza Mohammadi is *currently teaching at the university. I finished my Ms program two years ago and I am now I try to obtain a scholarship for a PhD.*

GENERALIZED EXTREME VALUE ADDITIVE MODELS VIA VARIATIONAL BAYES

Sarah E. Neville[1], M.J. Palmer[2], M.P. Wand[1]

[1]Centre for Statistical and Survey Methodology
School of Mathematics and Applied Statistics
University of Wollongong, Wollongong 2522, AUSTRALIA
sen045@uow.edu.au

[2] Commonwealth Scientific and Industrial Research Organisation
Mathematics, In formatics and Statistics Floreat 6014, AUSTRALIA

Analysis of sample extremes is becoming more prominent, largely driven by increasing interest in climate change research. In the past decade, additive models for sample extremes have been developed, ranging between Bayesian and non-Bayesian approaches, with various methods of fitting employed. We devise a variational Bayes algorithm for fast approximate inference in Generalised Extreme Value additive model analysis. Such models are useful for flexibility assessing the impact of continuous predictor variables on sample extremes. The crux of the methodology is variational Bayes inference for elaborate distributions. It builds on the work of Wand, Ormerod, Padoan & Fruhwirth (2010), in which the notion of auxiliary mixture sampling (e.g. Fruhwirth-Schnatter and Wagner, 2006) is used to handle troublesome response distributions such as GEV. The new methodology, whilst approximate, allows large Bayesian models to be fitted and assessed without the significant computing costs of Monte Carlo methods. A much faster analysis results, with little loss in accuracy. We include an illustration of the variational Bayes methodology using maximum rainfall data from Sydney and surrounding regions. This work has been done jointly with Mark Palmer (Commonwealth Scientific and Industrial Organisation) and Matt Wand (University of Wollongong).

References

Wand, Ormerod, Padoan and Fruhwirth (2010). Variational Bayes for Elaborate Distributions. Unpublished manuscript.

Fruhwirth-Schnatter and Wagner (2006). Auxiliary mixture sampling for parameter-driven models of time series of counts with applications to state space modelling. Biometrika, 93, 827-841.

Sarah Neville *is in her first year of a PhD with Matt Wand at the University of Wollongong. After graduating with honours from the same university in 2008, Sarah took part in the cadetship program at the Australian Bureau of Statistics in Canberra, followed by a stint at NSW Health in the Biostatistics training program in North Sydney. She is looking forward to an academic career with a balance of research and teaching.*

A DISTRIBUTION THEORY FOR TESTING THE EXISTENCE OF SPATIAL CLUSTERS

Yoshiyuki Ninomiya

Graduate School of Mathematics, Kyushu University; 744 Moto-oka, Nishi-ku, Fukuoka 819-0395, Japan; nino@math.kyushu-u.ac.jp

Testing the existence of spatial clusters is required in various fields including epidemiology, biology and astronomy. When the position and size of the cluster are unknown, the test can be regarded as a multiple testing problem, and then the p-value for the likelihood ratio test becomes asymptotically an exceedance probability for a discretely sampled Gaussian random field. One of the easiest ways for the evaluation of the exceedance probability is to use Bonferroni's method. It provides conservative testing, but it tends to be too conservative because usually there are high correlations among the multiple tests. On the other hand, several geometrical approaches have been recently proposed to evaluate an upper tight bound of the exceedance probability for a discretely sampled random field. In this presentation, such approaches are developed for our testing problem, and it is shown to be valid through simulation studies.

References

Taylor, J. E., Worsley, K. J. and Gosselin, F.
(2007). Maxima of discretely sampled random fields, with an application to ‘bubbles’. Biometrika, 94, 1-18.

Ninomiya, Y. and Fujisawa, H.
(2007). A conservative test for multiple comparison based on highly correlated test statistics. Biometrics, 63, 1135-1142.

Yoshiyuki Ninomiya *currently works as an associate professor in the Graduate School of Mathematics, Kyushu University, Japan. His area of interest is model selection for irregular statistical models including the change-point model and the exploratory factor analysis model.*

ASYMPTOTIC EXPANSIONS OF THE DISTRIBUTIONS OF THE ADF CHI-SQUARE STATISTIC IN COVARIANCE STRUCTURES

Haruhiko Ogasawara

Otaru University of Commerce
3-5-21 Midori, Otaru 047-8501 Japan
hogasa@res.otaru-uc.ac.jp

An asymptotic expansion of the null distribution of the chi-square statistic based on the asymptotically distribution-free theory for general covariance structures is derived under nonnormality. The added higher-order term in the approximate density is given by a weighted sum of chi-square distributed variables with different degrees of freedom. A formula for the corresponding Bartlett correction is also shown without using the above asymptotic expansion. Under a fixed alternative hypothesis, the Edgeworth expansion of the distribution of the standardized chi-square statistic is given up to order O(1/n). From the intermediate results of the asymptotic expansions for the chi-square statistics, asymptotic expansions of the joint distributions of the parameter estimators both under the null and fixed alternative hypotheses are derived up to order O(1/n). For detailed results corresponding to this abstract, see Ogasawara (2009, 2010).

References

Ogasawara, H. (2009). Asymptotic expansions of the distributions of the chi-square statistic based on the asymptotically distribution-free theory in covariance structures. Journal of Statistical Planning and Inference, 139, 3246-3261.

Ogasawara, H. (2010). Supplement to the paper “Asymptotic expansions of the distributions of the chi-square statistic based on the asymptotically distribution-free theory in covariance structures”. Economic Review (Otaru University of Commerce), 60 (4), 187-200. Retrieved from http://www.res.otaru-uc.ac.jp /~hogasa/

Haruhiko Ogasawara *is Professor of Statistics, Department of Information and Management Science, Otaru University of Commerce. He graduated from the Department of Educational Psychology, University of Tokyo, and obtained his Ph.D. degree at Tokyo Institute of Technology. His current area of interest is asymptotic expansions of statistics used in the behavioral sciences.*

HOLT'S LINEAR EXPONENTIAL SMOOTHING METHOD AUGMENTED WITH REGRESSORS

Ahmad Farid Osman[1], Maxwell L. King[2]

[1] Postgraduate Student, Department of Econometrics and Business Statistics
Monash University, Vic 3800, Australia
ahmad.osman@buseco.monash.edu.au

[2] Pro Vice-Chancellor, Research and Research Training, Monash University, Vic 3800, Australia.
Max.King@adm.monash.edu.au

Holt's linear exponential smoothing method was introduced in 1957 as an extended version of the single exponential smoothing method. The idea of this approach is to generate forecasts based on two time series elements, namely level and trend. In this paper, we adopt a similar approach in order to integrate the existing exponential smoothing method with regressors whose coefficients are time varying. We then translate this model into an equivalent state space structure which allows all the parameters to be estimated via the maximum likelihood estimation procedure. To avoid an explosion in the regression coefficient's value, we propose a slight modification that puts the updating process on hold until sufficient information on how the coefficient might have changed is available. We finally test the new forecasting approach using some simulated series and compare its forecast performance with that of traditional exponential smoothing models and other forecasting approaches that allow the integration of regressors into the model.

References

Holt, C.C. (1957). Forecasting Seasonal and Trends by Exponentially Weighted Moving Averages. Office of Naval Research, Research Memorandum No. 52.

Hyndman, R. J., Koehler, A. B., Ord, J. K. & Snyder, R. D. (2002). A State Space Framework for Automatic Forecasting Using Exponential Smoothing Methods. International Journal of Forecasting, 18, 439-454.

Hyndman, R. J., Koehler, A. B., Ord, J. K. & Snyder, R. D. (2008). Forecasting with Exponential Smoothing: The State Space Approach. Berlin: Springer-Verlag.

Ahmad Farid Osman *is a postgraduate student at the Department of Econometrics and Business Statistics, Monash University, Clayton campus, Victoria. He is currently pursuing his PhD study under the academic training scheme which is jointly sponsored by the Ministry of Higher Education, Malaysia and the University of Malaya, Kuala Lumpur, Malaysia. His main research interest areas are in time series analysis and econometrics modelling.*

APPLICATION OF GENERALIZED LINEAR MIXED MODELS TO GRAZING SYSTEMS

David Reid[1], Trevor Hall[2], John McIvor[3]

[1] Agri-Science Queensland, DEEDI
PO Box 6014, Red Hill, Rockhampton Qld 4701
David.Reid@deedi.qld.gov.au

[2] Agri-Science Queensland, DEEDI
PO Box 102, Toowoomba Qld 4350
Trevor.Hall@deedi.qld.gov.au

[3] CSIRO
306 Carmody Rd, St Lucia, Brisbane Qld 4067
John.Mcivor@csiro.au

A four-year, producer-inspired research project, jointly-funded by DEEDI, CSIRO and MLA, investigated different grazing systems across the northern beef industry. Nine commercial beef properties, each with two or three planned grazing systems (continuous, rotational, and cell - increasing grazing system intensity) were selected in north and south Queensland on brigalow (heavy, higher fertility soils) and eucalypt (light, lower fertility soils) land types. From one to six paddocks per system were selected on each property (total of 74 paddocks) for soil and pasture measurements following the growing seasons of 2006, 2007 and 2009. Spatial analysis identified each paddock as spatially uniform (assigned 1) or non-uniform (assigned 0) with respect to each pasture measurement (e.g. grass basal area) with the aim of testing the hypothesis that, as grazing system intensity (number of paddocks, frequency of cattle movements, capital costs, management input, etc) increases, spatial uniformity of grazing increases. Analysis as a Generalised Linear Mixed Model appropriately accommodated the multi-strata design (land types and location at a property level and grazing system at a paddock level) and the binomial nature of the data.

David Reid *is currently employed as a biometrician with Agri-Science Queensland (a service of the Department of Employment, Economic Development and Innovation, formerly DPI&F) in Rockhampton. He has over 23 years experience as a consulting biometrician with experience in a range of agricultural industries including horticulture, field crops, beef cattle, fisheries and forestry. David has specific expertise in farming systems research, particularly in design and analysis of on-farm research trials. He also provides statistical support and advice to post-graduate students at the local university.*

SPATIO-TEMPORAL DISEASE MAPPING OF FOOT AND MOUTH DISEASE IN VIETNAM

Kate Richards[1], *Martin Hazelton*[2], *Nguyen van Long*[3], *Mark Stevenson*[4]

[1] *Massey University, New Zealand*
Private Bag 11222, Palmerston North, New Zealand
kkrichards@hotmail.com

[2] *Massey University, New Zealand*
M.Hazelton@massey.ac.nz

[3] *Massey University, New Zealand and DAH Vietnam*
nvlong@dah.gov.vn

[4] *Massey University, New Zealand*
M.Stevenson@massey.ac.nz

Foot and mouth disease is a virus that can be transmitted by direct and indirect animal contact as well as by airborne means. It can affect all types of cloven-hoofed animals, but is found principally in cattle, sheep and pigs. Foot and mouth disease worldwide is listed as one of the top 10 agricultural diseases, with an economic impact estimated to be well into the billions of dollars (US), through loss of stock, trade losses, vaccine costs etc. Foot and mouth disease is of critical importance in Vietnam where the main agricultural animals are buffalo, cows and pigs which are all susceptible to this disease. There is considerable animal migration present, creating the potential for devastating spread of the disease over the whole country.

In this poster we describe the application of modern disease mapping techniques to better understand the spatio-temporal distribution of foot and mouth disease in Vietnam, and in particular the patterns in the distribution of its three major serotypes. The available data comprise monthly disease counts by province from March 2006 to January 2009. The data for serotype are not complete, with variable rates of serotype testing present across the country. This provides problems in modelling the data, but also provides the opportunity for our model to be used in the prediction of local disease strains in areas where this information is not directly available. We employ Poisson log-linear models for the disease counts, incorporating conditional autoregressive processes in the linear predictor to account for spatial and temporal dependency in the data. We fit them using MCMC methods implemented in the WinBUGS package. Our models identify two areas of Vietnam with significantly elevated risk, and suggest an overall decrease in the risk of disease from 2006 to 2009. This type of information could inform schemes for vaccination distribution. At present these vaccines are rather expensive, requiring booster vaccine to maintain immunity.

Kate Richards *recently completed a Bachelor of Science, majoring in Statistics and Physiology in 2009. She is currently completing Honours in Statistics at Massey University, Palmerston North, New Zealand. Her area of research interest is statistics in veterinary epidemiology, where she is looking at mapping of infectious diseases.*

MODELLING AND ESTIMATION OF FINANCIAL DATA USING MULTIFRACTAL EMBEDDED BRANCHING PROCESSES

O. D. Jones[1], D. A. Rolls[2]

[1]Department of Mathematics and Statistics, University of Melbourne VIC 3010
odjones@unimelb.edu.au

[2] Department of Medicine - Royal Melbourne Hospital, University of Melbourne VIC 3010
drolls@unimelb.edu.au

The use of models based on time-changed Brownian motions $X = B \circ \theta$ has been proposed for applications in finance, turbulence and telecommunications where "volatility clustering" has been observed. Here we imagine B is a Brownian motion and θ is a continuous process, called a "chronometer". A Multifractal Embedded Branching Process (MEBP) is a stochastic model based on a recent concept called the "crossing tree" [Jones et al., 2004] which a builds a tree from the pattern of ups and downs taken by the process to first passages of the integers (in the spatial dimension). A key idea is that for a time-changed Brownian motion, the up/down pattern is unaffected by changes in the time-dimension. For a time-changed Brownian motion the crossing tree is a Galton-Watson process and the family size distribution is particularly easy to write down. For an MEBP, random weights are assigned to the branches of the crossing tree and times are formed by taking products down the lines of descent.

We describe efforts to fit MEBP processes to financial data (e.g., currency exchange rate data), and show how they are able to simply capture important features in returns such as tails heavier than Gaussian and short range dependence in the returns but long-range dependence in the squared returns [Heyde et al., 2001].

References

Heyde, C.C. and Liu S.S. (2001). Empirical realities for a minimal description risk asset model- the need for fractal features. Journal of the Korean Mathematical Society, 38, 1047-1059.
Jones, O. and Shen, Y. (2004). Estimating the Hurst index of a self-similar process via the crossing tree. Signal Processing Letters, 11, 416-419.

David Rolls *has been a Research Fellow at the University of Melbourne since 2007. He received his Ph.D. from Queen's University at Kingston, Canada in 2003 and worked in the United States before coming to Australia. His research interests generally involve stochastic modelling and simulation. His Ph.D. thesis and research immediately after involved models for data network traffic exhibiting long-range dependence and heavy-tailed distributions. Currently his research is in two areas: the modelling of the transmission of infectious diseases over social networks, and the modelling of financial data with multifractal embedded branching processes.*

STOCHASTIC TESTS OF ORTHOGONAL EQUIVALENCE OF 3-TENSORS

Sakata, T.[1], Maehara, K.[2], Sumi,T.[3] and Miyazaki, M.[4]

[1] Department of Human Science, Kyushu University
4-9-1, Shiobaru Minami-ku, Fukuoka, 815-8540, Japan
sakata@design.kyushu-u.ac.jp

[2] Graduate School of Design, Kyushu University
4-9-1, Shiobaru Minami-ku, Fukuoka, 815-8540, Japan
kazumits@gmail.com

[3] Department of Human Science, Kyushu University
4-9-1, Shiobaru Minami-ku, Fukuoka, 815-8540, Japan
sumi@design.kyushu-u.ac.jp

[4] Department of Mathematics, Kyoto University of Education
Fujinomori-cho, Fukakusa, Fushimi-ku Kyoto, 612-8522, Japan
g53448@kyokyo-u.ac.jp

Recently multi-array data $T=(A_1;A_2;...;A_p)$ with p slices of square matrices, called 3-tensors, have been successfully used in various applied fields. For 3-tensors, $p-$, $q-$ and $r-$ transformations (a multiplication of a nonsingular matrix P from left, Q from right, and a replacement of each slice by a linear sum of three slices with coefficients from a nonsingular matrix R) are defined and equivalence is defined as mutual reachability by a mixed sequence of these transformations. Under equivalence tensor rank, a concept of the complexity of data, is invariant, and so it is important. To show equivalence is quite difficult because we need to solve a system of algebraic equations with too many variables. Therefore, we consider equivalence through the window of the determinant polynomial of a tensor T, $f_T(\mathrm{X})=\det(\sum_{i=1}^{p}x_iA_i)$ (for example, see Sakata et al.(2009)). If two tensors T_1 and T_2 are equivalent, it holds that $f_{T_2}(\mathrm{X})=cf_{T_1}(\mathrm{X}R)$ for some constant c and a nonsingular matrix R. In fortunate cases this relation among determinant polynomials can be deterministically checked, however, it is difficult to check in general. So, in this paper, we propose a stochastic technique of testing equivalence when P, Q and R are restricted to orthogonal matrices. This restriction is a natural one and considered as the real version of unitary equivalence of quantum states in quantum communication theory. The stochastic test is performed by discriminating the distributions of the determinant polynomial $f_T(\mathrm{X}R)$ where R is a random orthogonal matrix obeying the Haar measure on the orthogonal group *O(p).* If two tensors are orthogonally equivalent their distributions are the same and this is checked by looking the histograms or Kormogorov-Smirnov test. Further the moments of two polynomials can be compared by exact calculations, based on Weingarten function argued in Collins and Matsumoto (2009).

References

Collins, B. and Matsumoto, S. (2009). On some properties of orthogonal Weingarten functions. *arXiv:0903.5143v1*.

Sakata, T., Sumi T. and Miyazaki, M. (2009). Exceptional tensors with three slices and the positivity of its determinant polynomial. Abstract book, 57-th I.S.I conference, CMP37, p349.

Sumi, T. Miyazaki, M. and Sakata, T. (2010). About the maximal rank of 3-tensors over the real and the complex number field. To appear in Annals of Institute of Statistical Mathematics, Special edition of Älgebraic Methods in Computational Statistics".

Toshio Sakata *works as a professor of statistics in the department of Human Science of Faculty of Design of Kyushu University in Japan. His area of interest is mathematical problems and applications in tensor data analysis and multivariate analysis in ergonometric data and design data. Prof Toshio Sakata leads a tensor data analysis team in which coauthors, Prof. Toshio Sumi of Kyushu University, Prof. Mitsuhiro Miyazaki of Kyoto University of Education, and Kazumitsu Maehara, a p.h.d. student of Prof. Toshio Sakata are working together.*

ESTIMATING PARAMETERS IN QUERMASS-INTERACTION PROCESS

David Dereudre[1], Frédéric Lavancier[2], Katerina Stankova Helisová[3]

[1] University of Valenciennes, Laboratory of Mathematics and its Applications
le Mont Houy, 59313 Valenciennes 9, France
David.Dereudre@univ-valenciennes.fr

[2] University of Nantes, Faculty of Sciences et Techniques, Laboratory of Jean Leray
2 rue de la Houssiniere, BP 92208, 44322 Nantes 3, France
frederic.lavancier@univ-nantes.fr

[3] Czech Technical University in Prague, Faculty of Electrical Engineering
Department of Mathematics, Technická 2, 16627 Prague 6 - Dejvice, Czech republic
helisova@math.feld.cvut.cz

Consider a random set observed in a window $W \subset R^2$. The set is given by a union of interacting discs with randomly scattered centers and with arbitrary (random or deterministic) radii. Assume that its probability measure is given by a density with respect to the probability measure of a stationary Boolean model, i.e. with respect to a process of discs without any interactions, whose centers form a stationary Poisson point process with an intensity ρ . Next, assume that the density is of the form

$$f_\theta(\gamma) = \frac{e^{-H(\gamma)}}{Z_\theta} = \frac{e^{-(\theta_1 A(\bar{\gamma}) + \theta_2 L(\bar{\gamma}) + \theta_3 \chi(\bar{\gamma}))}}{Z_\theta}$$

for any finite configuration of discs γ , whose energy $H(\gamma)$ depends on the area $A(\bar{\gamma})$, the perimeter $L(\bar{\gamma})$ and the Euler-Poincare characteristic $\chi(\bar{\gamma})$ of the union $\bar{\gamma}$ composed of the discs from the configuration γ . Further, $(\theta_1, \theta_2, \theta_3)$ is a vector of parameters and Z_θ denotes a normalizing constant. Such a model is called Quermass-interaction process.

In this contribution, we describe a method for estimating the parameters $\theta_1, \theta_2, \theta_3$ and ρ studied in Dereudre et al(2010) which is based on Takacs-Fiksel procedure, and compare it with MCMC maximum likelihood method described in Møller and Helisová (2009).

References

Dereudre, D., Lavancier, F., and Helisová, K. (2010). Estimating parameters in Quermass-interaction process. In preparation.
Møller, J. and Helisová, K. (2009). Likelihood inference for unions of interacting discs. Scandinavian Journal of Statistics. Accepted.

Katerina Stankova Helisova *finished her PhD. study at the Department of Probability and Mathematical Statistics at Charles University in Prague in September 2009. She currently works as an assistant professor at the Czech Technical University in Prague, where she provides lectures of economic mathematics and exercises of basic mathematical analysis, and practices research. The area of her research is stochastic geometry and spatial statistics.*

ANALYSIS OF THE SIMILARITY MEASURES BETWEEN MOTIFS

Emi Tanaka[1] *and Uri Keich*[2]

[1] *PhD Research Student, Sydney University*
E.Tanaka@maths.usyd.edu.au

[2] *Senior Lecturer, Sydney University*
uri.keich@sydney.edu.au

Transcription factors regulate gene expression by binding to specific sites along the genome. The collection of these binding sites is referred to as a motif. Scientists working on gene regulation are often faced with the question of whether a motif they identified in a set of co-regulated sequences is significantly similar to any previously characterized motif. Habib et al (2008) claimed that a new motif similarity measure they developed, called BLiC, outperforms previous similarity measures. The BLiC score takes into account the similarity of the query motif to the database motif as well as their dissimilarity to the background distribution. This feature penalizes degenerate columns in the motif and helps reducing false hits in the database. We demonstrate, however, that the BLiC score exhibits characteristics which are highly undesirable for a motif similarity score. Instead we introduce an alternative method to penalize degenerate columns. Our novel approach offers comparable results with existing method

References

Gupta, S., Stamatoyannopoulos, J., Bailey, T., and Noble, W. (2007). Quantifying similarity between motifs. Genome Biology 8 (2).
Habib, N., Kaplan, T., Margalit, H., and Friedman, N. (2008). A novel Bayesian DNA motif comparison method for clustering and retrieval. PLoS computational biology, 4 (2).

Emi Tanaka *is currently a PhD research student in School of Mathematics and Statistics, Sydney University. Her area of interest is the application of statistics to genomics, in particular analyzing similarity measures between motifs.*

EFFICIENCY OF THE SCOTT-WILD ESTIMATOR FOR GENERALISED CASE-CONTROL STUDIES UNDER GENERAL MISSPECIFICATION

Jennifer Wilcock[1], *Alan Lee*[2]

[1] *Department of Statistics, University of Auckland*
Private Bag 92019, Auckland 1142, New Zealand
j.wilcock@auckland.ac.nz

[2] *Department of Statistics, University of Auckland*
Private Bag 92019, Auckland 1142, New Zealand
aj.lee@auckland.ac.nz

Data collected by response-selective sampling, where the probability of selection depends on the value of the response variable, requires different methods of analysis than data collected prospectively. Typical examples of studies where response-selective sampling is used are two- and multi-phase case-control studies, where data on some variables is collected using sampling rates that depend on the value of the variables observed at the first stage. Alastair Scott and Chris Wild (1997) have introduced a profile likelihood approach to estimation in this context, and various ad-hoc approaches to proving the efficiency of these methods have been proposed. In this talk, we describe a simple and general approach, based on projections and an adaptation of the method of Newey (1994), to demonstrating the efficiency of these methods under general misspecification.

References

Newey, W.K. (1994) 'The asymptotic variance of semiparametric estimators'. Econometrica, 62:1349-1382.

Scott, A.J. & Wild, C.J. (1997) 'Fitting regression models to case-control data by maximum likelihood'. Biometrika, 84:57-71.

Jennifer Wilcock *is currently a PhD student in statistics working with Professor Alan Lee at the University of Auckland. She has a professional background in civil engineering and public policy analysis and has worked in the fields of transport planning and engineering in both consultancy and local government environments in New Zealand and the UK. She specialised in transport modelling work, mainly network optimisation models and modal choice models for forecasting public transport use. She returned to university to study statistics to learn more about the theory underpinning statistical modelling.*

SWITCHING OFF ELECTRIC ANTS

Carole Wright[1], Gary Morton[2], Donna Baldwin[3]

[1] Agri-Science Queensland
180-202 River Boulevard, Oonoonba, QLD 4811
Carole.Wright@deedi.qld.gov.au

[2] Biosecurity Queensland
21-23 Redden Street, Cairns, QLD 4870
Gary.J.Morton@deedi.qld.gov.au

[3] Biosecurity Queensland
21-23 Redden Street, Cairns, QLD 4870
Donna.Baldwin@deedi.qld.gov.au

Electric ants (Wasimannia auropunctata) were first identified in Cairns, Far North Queensland in May 2006. They are an exotic environmental pest and listed as one of the world's top 100 invasive species. Although just 1-1.5mm long their sting causes a painful and itchy red welt. If established in Australia electric ants could become a significant agricultural pest by negatively impacting on native ant and insect populations and therefore indirectly increasing the occurrence of viral and fungal infections.

In an effort to eradicate electric ants from the Far North, Biosecurity Queensland is undertaking an extensive surveillance and baiting scheme. To maximise the effectiveness of the baiting strategy a monitoring experiment was performed to investigate the activity of electric ants over a 24 hour period. The experiment was conducted during both the cooler dry season (July) and twice in the hot and humid wet season (November, December). Knowing the behaviour of electric ants could dramatically increase the likelihood and number of ants coming into contact with the bait stations, reduce resource requirements and therefore cut monitoring costs.

Fitted splines showed that electric ants moved faster during the hotter parts of the day and that their activity was highly correlated with temperature and humidity. Electric ants monitored on feeding lines in the open, such as on pavements and fallen logs, showed a much stronger correlation to temperature and humidity than feeding lines in more protected places, such as tree branches and trunks.

These results have helped Biosecurity Queensland understand electric ant activity and therefore improve their surveillance and baiting schemes in the hope of eradicating electric ants from the Far North.

Carole Wright *is currently employed as a Biometrician with Agri-Science Queensland in Townsville. She is involved as a consultant on a large number of research projects across Far North Queensland, predominately in the areas of tropical horticulture and forestry.*

VARIANCE COMPONENTS ANALYSIS OF AN ORDERED CATEGORICAL OUTCOME IN NUCLEAR FAMILIES USING MARKOV CHAIN MONTE CARLO

Sophie G. Zaloumis[1], Lyle C. Gurrin[2], Stephen B. Harrap[1], Katrina J. Scurrah[2]

[1]Department of Physiology
University of Melbourne
Parkville, VIC 3010, Australia
s.zaloumis@pgrad.unimelb.edu.au
s.harrap@unimelb.edu.au

[2]Centre for Molecular, Environmental, Genetic & Analytic Epidemiology
Melbourne School of Population Health 1/723 Swanston Street
Carlton, Victoria, 3010, Australia
kscurrah@unimelb.edu.au
lgurrin@unimelb.edu.au

Statistical methods for the analysis of data from family-based genetic association studies are well-developed for continuously valued and binary outcome measures. Ordinal categorical outcomes have received much less attention, and are often analysed as unordered categorical responses or simply reduced to binary measures.

Methods to examine whether genetic and/or environmental sources can account for the residual variation in ordinal family data are currently not available. Such models usually assume proportional odds, that is, that the increased risk of realizing the higher of two adjacent outcome categories conferred by a unit increase in exposure is constant across outcome categories. The motivation for this assumption is to simplify the analysis and it is rarely investigated in practice.

These difficulties can be overcome by taking a simulation-based approach to model fitting using Markov Chain Monte Carlo (MCMC) methods. This approach has been used to fit a newly developed model (that allows the proportional odds assumption to be relaxed) to analyze the contribution of shared genetic effects and common family environment to the variation in Male Pattern Baldness data from the Victorian Family Heart Study. An association between MPB (as a four-category ordinal outcome variable) and a Single Nucleotide Polymorphism (rs6152, in the StuI gene, one of the androgen receptor family) was found. The effect of rs6152 depends on baldness stage. Carriers of the G allele have a lower threshold (logit scale) of vertex balding [κstui2: -0.37, 95% CI: (-0.68, -0.02)], which reduces a carrier's probability of frontal balding but increases their probability of vertex balding. Carriers of the G allele also have a higher threshold of both frontal and vertex balding [κstui3: 0.43, 95% CI: (0.04, 0.79)], and consequently, their probability of vertex balding is further increased, while their probability of hairloss in both the frontal and vertex regions is reduced.

Sophie Zaloumis *is a PhD candidate in the Department of Physiology, University of Melbourne. In 2006 she began working in the Department of Physiology as a Research Assistant. Her work involved examining the correlations and variance components of blood pressure phenotypes, and developing statistical models for ordinal categorical family data.*

OZCOTS 2010

7th Australian Conference on Teaching Statistics

The theme of OZCOTS 2010 is **Building capacity in statistics education.**

Statistical education is of vital significance to all statisticians and the statistical profession across the spectrum of educational levels and disciplines. The supply of statisticians is just one aspect of this spectrum, and the education of future consumers, users, producers, developers and *researchers* of statistics is both challenging and important for a modern information society and hence to the statistical profession. OZCOTS 2010 is of interest to all involved in the teaching and learning of statistics, including universities, colleges and schools, industries and governments.

The OZCOTS program includes keynote and contributed papers and forum discussion on topics across the statistical education spectrum of interest to statisticians and the statistical profession. Like all the conferences of the International Association for Statistical Education's (IASE), an optional refereeing process is offered to authors with papers accepted as refereed designated as such in the proceedings. The OZCOTS 2010 Proceedings are published on http://www.oznznets.com/ozcots2010.html

OZCOTS background
The first OZCOTS was run in 1998 by Brian Phillips with papers by the Australian speakers from the 5th International Conference on Teaching Statistics (ICOTS) which had been held in Singapore earlier in 1998. Its success in bringing together Australians involved in teaching statistics resulted in Brian and his Melbourne colleagues organising annual OZCOTS gatherings from 1999 to 2002. In 2006 Helen MacGillivray was awarded one of the first Australian Learning and Teaching Council's Senior Fellowships, with her fellowship programme to run throughout 2008. As part of her fellowship programme, Helen revived OZCOTS with Brian's help, and ran it as a two-day satellite to the 2008 Australian Statistical Conference (ASC), with a one-day overlap open to all ASC delegates who could also choose to register for the second day. The OZCOTS 2008 invited speakers were all funded as part of Helen's fellowship. OZCOTS 2008 was modelled on International Association for Statistical Education's (IASE) conferences, with papers in proceedings and an optional refereeing process offered to authors. The success of OZCOTS 2008 has lead to OZCOTS 2010.

OZCOTS 2010 gratefully acknowledges the support of the Australian Statistical Conference 2010, SAS and the NSW Branch of SSAI.

OZCOTS 2010 Conference Committee

Helen MacGillivray (joint chair, joint editor), Queensland University of Technology
Brian Phillips (joint chair, joint editor), Swinburne University
Alexandra Bremner (local arrangements), University of Western Australia

OZCOTS 2010 gratefully acknowledges the following referees for their assistance: Michael Bulmer, Rosemary Callingham, Mike Forster, Glenda Francis, Ian Gordon, Eric Sowey, Christine McDonald, Katie Makar, Michael Martin, Peter Martin, Peter Petocz, Matt Regan, Jackie Reid, Richard Wilson, Therese Wilson.

KEYNOTE SPEAKER

WHAT I SEE IS NOT QUITE THE WAY IT REALLY IS

Chris J Wild

Department of Statistics, University of Auckland
38 Princes St, Auckland, New Zealand
c.wild@auckland.ac.nz

In this talk we will gaze through the ripple glass of a bathroom window and wander Alice-like down garden paths through a wonderland where what we see is never quite the way it really is. The paths our odyssey leads us along are conceptual pathways that start with conceptualisations of statistical inference that are intended to be accessible to, and operable by, students mid-way through high school and lead us, via a series of connected trails, all the way to plot annotations that better reveal the stories being told by factor variables in generalised linear models. Along the way, both motivating and suggesting ways forward for all of this, we will meet novel visualisations of sampling variation, re-sampling variation and randomisation variation. The talk will draw on a paper with Maxine Pfannkuch, Matt Regan and Nicholas Horton entitled, “Towards more accessible conceptions of statistical inference” to be read to the Royal Statistical Society late in 2010 and on other work on making inference more accessible, particularly via visualisations, with these and other collaborators.

Chris Wild *- Professor of Statistics at the University of Auckland, New Zealand and recognised by Fellowships of the American Statistical Association and the Royal Society of New Zealand, Chris Wild is a member of a rare crossover species. He publishes extensively in statistical methodology, particularly on response-selective and missing data problems, but also works substantively in statistics education. He co-wrote the Wiley books Nonlinear Regression (1989) and Chance Encounters (2000) with George Seber. His best known statistics education paper is Statistical Thinking in Empirical Enquiry with Maxine Pfannkuch (1999, International Statistical Review). Chris' interests in statistics education include curricular revolution at school levels, growing university statistics programmes, and improving the penetration, quality and practical impact of statistics education at all levels. Chris has been a Council member of the International Statistical Institute, President of the International Association for Statistics Education and an Associate Editor of the International Statistical Review, Biometrics, the Statistics Education Research Journal, and ANZJS. He was Head of Auckland's Department of Statistics 2003-2007 and co-led the University of Auckland's first-year statistics teaching team to a national teaching award in 2003. His keynote addresses include the Royal Statistical Society, the Statistical Society of Canada, and ICOTS.*

STATISTICAL CONSULTING WITH POST-GRADUATE STUDENTS

FINCH, Sue and GORDON, Ian

University of Melbourne, Australia
Sfinch@unimelb.edu.au

The Statistical Consulting Centre at The University of Melbourne provides a post-graduate consulting service to higher degree research students across all disciplines from medicine to traditionally non-quantitative fields like development studies. Students can obtain advice about any stage of the research cycle from refining their research question and developing a suitable design to presenting and communicating findings. Some students have very little knowledge or experience in statistics, except in collecting their research data. Others are competent but need particular advice about specific problems. We took the opportunity to survey consulting sessions with post-graduate student, in order to reflect on the educational needs of these clients. We report on our survey and present some case studies to characterise aspects of graduate education that are applied and contextually relevant.

A MODEL FOR BUILDING STATISTICAL CAPACITY AMONG SCIENCE RESEARCH STUDENTS

BISHOP, Glenys and WILLIAMS, Emlyn

Statistical Consulting Unit, The Australian National University, Australia
glenys.bishop@anu.edu.au

The Statistical Consulting Unit at the Australian National University provides support to Honours and graduate research students, in addition to research staff. Students may consult one of the statisticians in the unit about design of the data collection phase such as design of an experiment or survey. They may also seek assistance with analysis, to the point where they are sufficiently adept to apply the recommended techniques themselves. In addition to the benefits of a better thesis outcome, students have also learnt the value of collaboration and been exposed to a range of design and analysis possibilities that may not have been available to them before consulting the unit.

However, to make the consulting process more efficient, it is important that students have enough statistical knowledge to be able to communicate effectively with statistical consultants. The Statistical Consulting Unit, at the request of the Colleges of Science, established a project to investigate the most practical and effective ways to transfer knowledge about statistical practice and statistical thinking to ANU Science students. After extensive consultation with stakeholders, a number of options, including their advantages and disadvantages, were proposed. An online learning model was adopted as the main method to build statistical capacity among Science research students. This can be supplemented by short courses addressing specific needs of some groups within the university.

This paper will report on a number of issues which had to be considered in arriving at this model, including the availability of existing online courses, the experiences of other institutions, whether to make an online course compulsory or voluntary, how to encourage students to take a voluntary online course, how many and which topics to include, ease of implementation and constraints of the university environment. These will be illustrated where possible using examples from the implementation.

TRAINING FOR STATISTICAL COMMUNICATION IN THE WORKPLACE

GIBBONS, Kristen[1] *and MacGILLIVRAY, Helen*[2]

[1] *Clinical Research Support Unit, Mater Medical Research Institute, Australia*
kgibbons@mmri.mater.org.au

[2] *Queensland University of Technology, Australia*

Transitioning from the role of a university student to a statistician working as a collaborative researcher, consultant and educator of doctors, nurses, scientists and allied health staff in a medical research environment is a very daunting, and challenging task, particularly in explaining about the use and misuse of statistics. While a sound and substantial undergraduate statistics background provides the necessary foundation for ongoing learning, and to competently understand and perform statistical analyses, it does not necessarily enable you to effectively communicate the concepts and results of these analyses to busy professionals with limited, and often no, prior statistical experience. The training of statisticians for such roles must incorporate the skills required for not only performing statistical analyses, but also for consulting with researchers with limited statistical knowledge, effectively teaching researchers both the concepts and the use of statistical software, as well as being able to ensure that the correct statistical advice is not lost in the power struggle of competitive research.

There has been increasing emphasis in statistics education on inclusion of experiential learning of the whole process of the statistical data investigation cycle, and on communicating statistics. As an undergraduate, the author was fortunate to have these learning experiences, but a further facet that has received little attention in statistics education has proved to be significantly advantageous. The experience gained by the author during their undergraduate degree in a developmental and mentored programme in tutoring statistics has proven to be invaluable in communicating with staff in a large tertiary hospital that also incorporates a basic science medical research institute. Through tutoring different aspects of statistics, including experiential learning of data investigations and problem-solving, as well as different disciplines of students, the principles of communicating statistics were learnt, and developed over a number of years. Without these tools the role of the statistician in a workplace where the majority of staff have a minimal statistics background would be near impossible.

DATA VISUALISATION: A NEW STATISTICAL LITERACY TOOL FOR STATISTICAL OFFICES

FORBES, Sharleen

Statistics New Zealand, Victoria University, New Zealand
sharleen.forbes@stats.govt.nz

The ability to harness massively increased computing power together with the availability of new free graphical tools easy to download from the Internet has led to a burst of creativity in new dynamic and interactive data visualisations. Some of this activity has taken place in national statistics offices. Official statistics have traditionally been released in simple and standard tables and graphs but these statistics provide the evidence base for much of government policy and new static and dynamic graphs and maps that allow users to interrogate and interact with data in new ways have been developed to increase the usefulness and understanding of these statistics. In this presentation a Consumers Price Index kaleidoscope is used to investigate the structure of index numbers, multidimensional scatterplots to teach conceptual understanding of multiple regression, and dynamic population pyramids, commuter flows and integrated graphs and maps to show the multi-disciplinary nature of statistics in the real world.

COMPARISONS BETWEEN MARKETING AND PSYCHOLOGY STUDENTS IN LEARNING STATISTICS

FRANCIS, Glenda and LIPSON, Kay

Faculty of Higher Education Lilydale, Swinburne University of Technology, Australia
klipson@swin.edu.au

This study investigates how the attitudes of marketing and psychology students towards statistics differ and whether the differences are inherent from the start or develop as they progress through their degree. The attitudes towards statistics for final year marketing and psychology students were measured using a slightly reduced version of Schau's 36 item SATS scale (Schau, 2005). The third year marketing students were shown to have much less positive attitudes towards statistics than their psychology counterparts, and perceive statistics as less useful to their discipline. Marketing students see only modest value in statistics at the start of their program, and unfortunately this worsens over time, while psychology students see statistics as more valuable from the start, and increase this view over their program. Both groups of students start out with almost the same level of interest in statistics but again this reduces markedly for the marketing students while level of interest increases for the psychology students. Psychology students' perception of the level of difficulty of statistics stays relatively constant during their program. While Marketing students see statistics as less difficult at the outset than psychology students, for them statistics becomes more difficult over their course of study. It is suggested that this difference may be as a result of differences in the two course structures and that embedding statistics more fully into specific discipline areas improves student' attitudes and helps to prepare them better for the workplace.

REALSTAT: FROM IDEAS TO DATA AND BEYOND

GORDON, Ian and FINCH, Sue

University of Melbourne, Australia
irg@unimelb.edu.au

RealStat is a suite of multimedia case studies in statistics, designed to be examined in a web browser and used in mainstream statistics teaching. In this paper we describe the educational philosophy behind RealStat, outline how the case studies have been developed, and present several of them. The case studies have been used at every level of courses at the University of Melbourne, from first year to Masters level. They are designed with levels of complexity and a variety of relevant statistical techniques and applications. Throughout, the aim is to inspire students of statistics with rich context and material, including visual media, graphics, context background, commentary by a statistical analyst and so on. The data sets themselves are available and the students can interact with the data and carry out their own analyses.

RealStat has been used in lectures, assignments and even examinations. We describe some of these uses, and the experiences of lecturers in developing curriculum material based on RealStat.

All case studies used have arisen through projects seen in the Statistical Consulting Centre at the University, and therefore have a history that is primarily local and current, enhancing their appeal.

SPREADSHEETS AND SIMULATION – A NEW WAY FORWARD FOR TEACHING STATISTICS

BARR, Graham and SCOTT, Leanne

Department of Statistical Sciences, University of Cape Town, South Africa
graham.barr@uct.ac.za

Fundamental statistical concepts remain elusive to many students in their introductory course on statistics. There are two teaching imperatives for educating statisticians, the one a more philosophical thrust to develop an appreciation of the nature of chance and the impact of random variation on our lives; and the other, the provision of a set of practical tools to support quantitative decision making. Becoming adept at the latter has often dominated introductory statistical courses, necessitating students investing large amounts of time in performing calculations to implement statistical tests. The advent of calculators with the ability to perform many of the standard statistical tests lead to a view that students could bypass the calculations, and that teachers could, instead, focus their efforts on the interpretation of test results. The experience of the authors has been that this approach, especially when coupled with poor educational backgrounds, has lead to students having very poor understanding of the rationale surrounding statistical tests. It is argued that the calculator approach to teaching appears, in reality, to undermine the development of an appreciation of fundamental statistical concepts. This talk will focus on the teaching of statistics within a spreadsheet environment, wherein the students are required to master the basics of Excel to perform statistical calculations. This approach has the advantages of developing the students' ability to work with data whilst also building an understanding of the elements embedded in the formulae which they use. At the same time teaching sessions are built around a suite of Excel based simulations which attempt to demonstrate the concept of random variation and show how statistical tools can be used to manage uncertainty. The authors' experiences with this teaching approach in an introductory statistics course involving some 1200 first year students in South Africa will form the basis of this paper.

REVISITING THE MISUSED, MISUNDERSTOOD AND UNLOVED STEM AND LEAF PLOT

GRIFFITHS, David[1] and STIRLING, Doug[2]

[1]University of Wollongong, Australia
griffd@uow.edu.au

[2]Massey University, New Zealand

When John Tukey unleashed some new tools and a new way of thinking about what until then were loosely categorised as descriptive statistics on the disciplinary stage nearly 40 years ago, it has been said that a new culture began.

This paper explores the effectiveness of that cultural change on how we teach and use techniques of exploratory data analysis, as well as embedding EDA into an earlier culture, when many of the EDA tools, including at least two of the supposedly new ones, were already in use.

The stem and leaf plot is one of those not so new tools. The principal foci of this paper are its history, its misuse and its underuse, as well as highlighting novel approaches to the use of stem and leaf plots.

What is its history? What are the catalysts for misuse and underuse? And what are the cures? The first answer is, unsurprisingly, that the history is not what you think. The second is easy to identify and multi-faceted, including materials (print and electronic) and resources (software), teachers, teaching and professional practice. The third may simply be recognition that the revolution that supposedly happened in the 1970's never did. It was merely a small change in mindset as to how we teach the first week of the first course in Statistics.

May this paper help to bring on the 2010 revolution!

INTRODUCTORY EPIDEMIOLOGY WITH A VIRTUAL POPULATION

BULMER, Michael[1] and KAPLAN, Daniel[2]

School of Mathematics and Physics, University of Queensland, Australia
m.bulmer@uq.edu.au[1]

Macalester College, USA[2]

We have developed an online population of virtual people that have been used as experimental subjects in student experiments for the last two years. However these virtual people also have genetic histories tracing back 240 years, records of disease and mortality, and a variety of demographic characteristics. We are now exploring this side of the virtual environment with students enrolled an introductory course on epidemiology. We will report on the types and outcomes of studies undertaken by students, their experiences in the environment and our reflections on the role of these experiences in learning.

NATIONAL STATISTICAL CURRICULUM JOURNEYS

MACGILLIVRAY, Helen

Queensland University of Technology, Australia
h.macgillivray@qut.edu.au

Since 1990, Statistics has featured in Australian school curricula. In the National Statement on Mathematics for Australian Schools (1991), Chance and Data was one of the five syllabi areas, with as many pages of the statement devoted to it as to Algebra. It is somewhat ironic that both these areas have been the source of much angst in the past two decades - of universal importance in each is development of their individual ways of *thinking*.

Reasons for, and sources of, frustrations in Chance and Data in school curricula are complex and should not be trivialized nor attributed narrowly or simplistically. They have been experienced not just by statisticians and statistics educators, but also by teachers, many of whom are eager to learn about, and teach, statistics in rich and engaging ways. It is informative and sobering to look at some history and timescales in the UK and USA, starting from the 1970's when Peter Holmes led the UK Schools Council Project on Statistical Education emphasizing that statistics should be taught, learnt and assessed in ways that reflect its thinking and practice, with real data and real experiences. This Project informed the American Statistical Association's Quantitative Literacy Project of the 1980s, and the Statistics Focus Group sponsored by the Mathematical Association of America's Curriculum Action Project in 1991 on developments for introductory tertiary courses. Their influences can be seen in the American Statistical Association's 2005 'Guidelines for assessment and instruction in statistics education' (GAISE) college report, and the 2007 GAISE pre-K-12 report (http://www.amstat.org/education/gaise/).

This paper gives a very brief overview of two decades of involvement in statistics education across school and transition levels. This has included designing and delivering many professional development workshops and enrichment programs, working with teachers and educational authorities, and, most recently, significant involvement in national curriculum challenges and writing materials for teachers. Pitfalls, progress, problems and promising avenues for ways forward are discussed. World-leading innovations and developments in statistics education have come from Australia and New Zealand, but there is much to be done in the national context.

KEYNOTE SPEAKER

TRANSFORMING STATISTICS EDUCATION IN SOUTH AFRICA

NORTH, Delia

University of KwaZulu-Natal, South Africa
northd@ukzn.ac.za

Challenges faced by statistics education in South Africa are magnifications of challenges everywhere for statistics – shortage of statistics educators, increasing need in society for statistical knowhow, education and professional development of teachers, increase in people coming to university from lower socio-economic backgrounds, mixture of cultures and multiple languages in one class room,....and many more!

The author attempts to come to grips with the current reality of statistics education in South Africa. In particular, the author focuses on the transformation of the South African education system, with respect to Mathematics and Statistics education. Major changes in Mathematics education at grass roots level has occurred over the last two decades as the government embarks on a process of grappling with legacies of the past, whilst balancing risks and opportunities of the future. Mathematics education is closely linked with statistics education, so that these changes have the potential to transform statistics education in South Africa.
A historical overview will be given, followed by a discussion of current challenges and successes that face statistics education at the various levels. The paper concludes with a discussion of various initiatives and proposals that have been undertaken to enhance statistics capacity in South Africa.

Though Statistics education in South Africa is relatively "young" compared to countries such as Australia, the problems faced are similar, with the result that lessons learnt by tackling South Africa's challenges might be useful to consider in statistics education in general.

Delia North *has over 25 years experience in teaching statistics across disciplines and levels in universities. Over the years she has become increasingly interested in statistics education, which resulted in her being appointed as the local chair of the 6th International Conference on Teaching Statistics [ICOTS6] which was held in Cape Town, South Africa, in July 2002. She has been a member of the Executive Committee of the South African Statistical Association [SASA] for more than ten years and has been chair of the Education Committee of SASA since 2003. Delia is very actively involved with voluntary work, conducting various outreach activities relating to the introduction of Statistics into the school syllabus for the first time ever in South Africa. She is known internationally for her work in supporting South African teachers and she is currently a vice-president of the International Association for Statistical Education [IASE].*

DYNAMIC DIAGRAMS FOR TEACHING DESIGN AND ANALYSIS OF EXPERIMENTS

STIRLING, Doug

Massey University, New Zealand
d.stirling@massey.ac.nz

Dynamic and interactive diagrams provide an effective way to teach introductory statistics through simulations and other illustrations of concepts. Many authors have written such diagrams, often in the form of Java applets.

Although there is similar potential in the use of applets to teach more advanced statistics, few have been published. One reason is that advanced statistical methods generally involve more complex numerical algorithms and are harder to illustrate graphically so dynamic diagrams are more difficult to program than those for basic statistical methods. The target audience of students who will benefit is also smaller, making the development overhead harder to justify.

This paper introduces a collection of over 200 applets for teaching the design and analysis of experiments that are contained in two CAST e-books (Stirling, 2010) about agricultural and industrial experiments. Some of these are 3-dimensional diagrams that can be rotated to show the properties of models for two or three factors. Others involve simulations to explain why some experimental designs are better than others. Many applets however were individually designed to illustrate specific concepts such as splitting the explained sums of squares in split plot experiments or randomisation of treatment allocation to experimental units.

There is also considerable potential in the use of dynamic diagrams to explain concepts to help teach other advanced statistical methods. A collection of CAST applets to teach multiple regression (Stirling, 2006) is briefly described and there is a short discussion about the potential for using dynamic diagrams to explain concepts in sample surveys and multivariate analysis.

Stirling, W. D. (2010), CAST release 4.1, http://cast.massey.ac.nz

Stirling, W. D. (2006), "Interactive 3-dimensional diagrams for teaching multiple regression", Proceedings of the Seventh International Conference on Teaching Statistics (2006).

A FREE STAND-ALONE SYSTEM FOR STATISTICAL DATA ANALYSIS WRITTEN IN R

CHANDRANANDA, Dineika and WILD, Chris

Department of Statistics, University of Auckland, New Zealand
c.wild@auckland.ac.nz

The authors are building a system for data analysis designed initially for use in New Zealand high schools but now with capabilities that will also make it useful in early-level university courses. Our desire has been to provide a tool that will actively encourage the exploration of multivariate data sets and enable emphasis to be kept almost entirely on seeing what data is saying rather than learning how to drive software. We want student-time primarily to be spent thinking about the questions they want to ask of the data and what the things they see mean rather than on the complexities and grind of getting hold of them. We also want to keep the data students are working with always “in their faces” to minimise abstraction. We have done these things using a drag and drop metaphor in which the software is driven by dragging the names of variables from the top of the data spreadsheet and dropping them in a small number of appropriate places. What is delivered instantly, and with an almost nonexistent learning curve, is graphics (plots). Numerical summaries and inferential information can also be obtained but the user has explicitly to ask for them. Plots involving up to three variables require almost no system knowledge. Gradually, as students get more experienced and the questions they want to ask become ambitious, they can learn more about system enabling them to dig in deeper. Relationships involving up to six variables can be explored using only very basic plot types. A useful set of tools for re-expressing variables is provided, but these are seen as existing for more experienced users. While the system is built in R, the user should never have to interact with R. We discuss pedagogical imperatives and describe system capabilities, design choices and the reasoning that led to them.

AN EXPERIMENTAL STUDY COMPARING STRATEGIES OF LEARNING HOW TO USE STATISTICAL SOFTWARE PACKAGES IN INTRODUCTORY STATISTICS COURSES

BAGLIN, James and DA COSTA, Cliff

School of Mathematics and Geospatial Sciences, RMIT University, Melbourne, Australia

In 2005, the American Statistics Association made the use of statistics packages, such as SPSS/PASW, Minitab, and R, a key recommendation in the Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report. Statistical software packages present instructors and students of introductory statistics courses with a number of benefits. These include automating difficult statistical formulae, offering instructors unique tools for demonstrating statistical concepts and giving students experience with packages that they may one day utilise in their careers. However, statistics instructors are conversant with some of the learning issues students encounter in using these packages. Anecdotally, most students demonstrate poor recall of learned material arising from week by week statistics computer tutorials and the poor recall is even more pronounced between semesters. As the student's proficiency with these packages is rarely formally assessed, little is known about efficient and effective ways of promoting this proficiency in the use of statistical packages. This study compared two strategies of learning how to use a statistical package: explicit instruction versus prompting. Explicit instruction gives students comprehensive written and visual instructions for operating a statistical package. Conversely, Prompting provides students with clues, but forces the student to largely discover how to operate the software package for themselves. Participants were recruited from a large introductory statistics course and randomly allocated to one of the learning strategies. The participants then worked through a single one-hour tutorial, based on their allocated learning strategy using the statistical package SPSS/PASW. At completion of the tutorial, the participants filled out a short questionnaire assessing their statistical package self-efficacy and anxiety levels. After one week, the participants were followed-up and given a short quiz testing their recall of the use of the statistical package from the first session. The primary analysis compared the difference in recall between the two learning strategies. In order to evaluate the learning context and potential trade-offs between learning strategies, the secondary analysis focused on comparing the differences in students' statistical package self-efficacy and anxiety. The major findings are discussed within the context of introductory statistics courses.

WWW MEANS WIN WIN WIN IN EDUCATION – SOME EXPERIENCES FROM ONLINE COURSES IN APPLIED STATISTICS

GELLERSTEDT, Martin and SVENSSON, Lars

Dept. of Economics and IT, University West Sweden, Sweden
martin.gellerstedt@hv.se

Challenge/objective: We wanted to offer an education in practical statistics arranged in a way that makes it possible to combine studies and work.

Pedagogical challenge: The University West has been offering distance education for one- and a half decade. Based on previous experiences and interviews with colleagues the following general factors were identified as the most crucial:

- Continuity
- Communication – "not being alone"
- Detailed instructions

Statistics – pedagogic challenge: It is rather common that statistics is regarded as boring, difficult and rather unimportant. Since we believe in the opposite we stated that the educational solution must show that statistics is fun, not that difficult and extremely important – not to say a necessity in the society of today.

Solution: We started the "SPSS academy" offering in total four courses (five week courses) in practical statistics using SPSS. All courses are completely web-based.

Practical solutions for taking account to the crucial factors:

- Combining theory and practical exercises in SPSS and part time examination gives continuity
- Intensive discussions on homepage, push-mails and quick responses
- More personal presentations and detailed study guidelines
- Intensive instructions for assignments
- Studio produced pre-recorded lectures

Evaluation: The SPSS academy has been running for one year now with full courses. The proportion of students completing a course is around 80%. Student evaluations shows that study guidelines, quick responses and pre-recorded lectures is highly appreciated. The University has reached new students; based on a questionnaire we estimate that more than 90% of the participants would not be able to follow an on campus course.

GENSTAT FOR TEACHING

BIGGS, Carey

VSN International, Hemel Hemstead, UK
Carey.Biggs@vsni.co.uk

VSNi believe that statistics and data analysis is central to many disciplines trying to bring meaning and deeper understanding to a student's research and work. A solid grounding in statistical practice is vital to many of today's researchers, as a part of this, students should be taught with tools they will encounter in later life, and tools that add value to their learning.

GenStat for Teaching is a set of new software tools, based on GenStat 13th edition. Currently a Schools edition is being developed for high school students of mathematics and statistics, and a university edition for undergraduates. Both systems retain the core statistical routines and strengths of GenStat, but have been specifically tailored to suit the needs of different level of students. Our developers have worked with high school teachers and university lecturers to provide a simpler menu based system highlighting the more usual statistics techniques that students come across; the more advanced university version unlocks slightly more detailed and complex statistics in line with the types of analysis techniques studied by undergraduates. At VSNi we are keen to support students and teachers alike, which is why GenStat for Teaching is free, at both schools and university level.

VSNi is a prime supplier of data analysis software for the biological and life sciences markets worldwide. We were formed in 2000 as a spin off from Rothamsted Research (RRES) and the Numerical Algorithms Group (NAG). We are backed by UK government through RRES being the largest land-based research institute in the UK and arguably the original home of statistics. As a result we are uniquely placed to provide statistical software to the world's educators and students. Our ethos is collaboration and partnership, ingrained within our psychology from our government links.

DIAGNOSTIC TESTING FOR STATISTICS COURSES

TOWNLEY-JONES, Maureen

School of Mathematical and Physical Sciences, University of Newcastle, Australia
Maureen.Townley-Jones@newcastle.edu.au

It is generally recognised that a part of any cohort of students entering tertiary education are underprepared in mathematical skills which are essential for understanding statistics at a tertiary level. As well as being underprepared, students are entering tertiary education with a misconception that mathematical and/or statistical skills are not required or misplaced perceptions of statistics due to learning experience in earlier school years prior to tertiary study. As tertiary educators, we recognise the importance of students having these skills to successfully navigate their way through university study. But students do not understand why basic mathematical and statistical skills are important or related to their discipline of study.

The consequences of allowing students to enrol in tertiary education with varying mathematical and statistical skills have been the basis for many Australian universities to implement various strategies that will allow students and staff to identify the weakness and strengths in students quantitative skills and direct students to other remedial actions, thereby to improve student retention and give students skills that are important graduate attributes.

This paper will discuss the journey that was initiated by the recognition of a need to do diagnostic testing for statistics and mathematics courses. The challenges of meeting different university student cohorts are discussed as well as strategies and the results of showing performance in first year subjects requiring mathematical skills compared to diagnostic test results.

THE IMPROVEMENT OF LEARNING PROCESS OF MATHEMATICAL STATISTICS I SUBJECT THROUGH STUDENT TEAMS ACHIEVEMENT DIVISIONS (STAD) METHOD USING ENGLISH AS MEDIUM LANGUAGE

SLAMET, Isnandar

Department of Mathematics, Sebelas Maret University, Surakarta, Indonesia
isnandar06@yahoo.com

The learning and teaching processes of Statistical Mathematics I subject in the Department of Mathematics, the University of Sebelas Maret Surakarta, Indonesia, used to follow a teacher-centered paradigm. Lecturers mostly dominated talk, delivered the material, gave examples and problems. Students recorded the material or made notes and solved the problems given. They are reluctant to ask questions even when they were asked to do so. It is important to note that the main competency of this subject is that students are expected to have knowledge, understanding, intellectual skills and critical thinking.

In Indonesia English is learnt as a foreign language and a compulsory subject in all levels of education. However, prior studies report that tertiary students face English language difficulties as a significant barrier to understanding material.

This study investigates the achievement of learning process of this subject given in the academic year 2007/2008 in the Department of Mathematics, the University of Sebelas Maret Surakarta, Indonesia through STAD (Student Teams Achievement Divisions) method using English as a passive medium language. To measure all these processes, assignments, quizzes, mid term test, group work, final test and questionnaire were given.

The result shows percentage of students passing this course is 98%. The study confirms the importance of student-centered learning process and the problems faced by students in the use of academic English.

‘SHE CAN’T DO SUMS A BIT’ OR CAN SHE?: TRACKING STUDENT LEARNING (CARROLL, 2002)

MORRIS, Maureen

University of Western Sydney, Australia
m.morris@uws.edu.au

This case study formed part of an action research project that was aimed at developing an aligned teaching/learning framework that effectively promoted the development of statistical thinking. It tracked changes in teaching and learning across five sessions in a foundation statistics subject at the University of Wollongong. Evaluation of the pedagogy engaged a mixed method approach. Evidence was derived from the multiple perspectives identified in student and teacher surveys, and assessment data. Grounded in experience and the literature, the teaching/learning framework included active participation in scaffolded learning tasks that centred on ‘real data’ .The intended learning was clear to the teacher, but did students embrace the processes and demonstrate the hoped for outcomes? Examination of assessment marks and student evaluation surveys detected student support for the pedagogy and positive changes in student results. However it was the detailed examination of the assessment tasks themselves that yielded evidence of a concurrent shift in cognitive demand. Increased teaching focus on the development of the higher order skills required for statistical thinking had not only enhanced achievement, but also promoted the framing of questions and marking criteria that detected demonstration of achievement.

A LEARNING DESIGN TO SUPPORT STUDENT LEARNING OF STATISTICS WITHIN AN ONLINE LEARNING ENVIRONMENT

BAHARUN, Norhayati[1] *and PORTER, Anne*[2]

[1]*University of Technology MARA (UiTM) Malaysia*
[2]*School of Mathematics and Applied Statistics*
University of Wollongong, Australia
nbb470@uow.edu.au

This paper presents the results of a study examining the use of a learning design map within an online learning environment in supporting the student learning of statistics at the University of Wollongong. This study compares two cohorts of undergraduate students who enrolled in an Introductory Statistics subject which were eighty-nine students in Autumn 2009 (before the implementation of learning design map within subject) and hundred eighty-four students in Autumn 2010. In Autumn 2010 session, this subject was designed based on the learning design representation (http://www.learningdesigns.uow.edu.au) which includes three major elements of learning activities such as tasks, resources, and supports. The aim of the learning design map is to provide guidance to students in learning this subject on a weekly basis through variety of learning resources made available via interactive links on e-Learning site which incorporates primary resources i.e. lecture notes, Edu-stream (audio recorded lectures); the tasks i.e. laboratory work, laboratory tests, worked solutions, data sets; other specific learning resources such as video resources to support most topics, SPSS notes; and ongoing support materials i.e. learning strategies, student forum, past exams and laboratory tests, student support advisers, learning modules, etc. The students were assessed on their basic statistical literacy and reasoning using a CAOS test (Comprehensive Assessment of Outcomes in Statistics from https://app.gen.umn.edu/artist/) at the beginning (pre-test) and the end of session (post-test) in Autumn 2010, on top of subject assessment i.e. laboratory tests and group project assignment. The findings show that the students in Autumn 2010 demonstrated improved performance from beginning to end of the session which were better than the students in Autumn 2009. Furthermore, at the end of session the students were asked to participate in an online survey and the results reveal that the learning design map implemented within subject e-Learning site has the potential to improve the student learning of statistics i.e. self-confident, engagement, peer interaction, reflection, etc which based on Boud and Prosser's framework of four principles of high quality learning. The paper concludes with a discussion on impact of learning designs on student learning of statistics and followed by suggestions for further research.

ENGAGING FIRST YEAR STUDENTS

KHAN, Nazim

School of Mathematics and Statistics, University of Western Australia, Australia
nazim@maths.uwa.edu.au

Most first year service units are difficult to teach as students are not interested and cannot see the relevance to their chosen major field of study. Additionally, students in such units are from a variety of backgrounds and mathematical ability, and are enrolled in very diverse courses. I will discuss one such unit, compulsory for all first years who come into Science with the lowest level mathematics qualification. Over the last few years this unit has had a negative perception amongst students and client faculties. I tell the story of how I approached this unit over two semesters. The student response, attitude and performance improved remarkably in these two semesters.

A CASE STUDY OF KNOWLEDGE OF KEY STATISTICAL CONCEPTS BEFORE AND AFTER AN INTRODUCTORY STATISTICS CLASS

JERSKY, Brian

Department of Statistics, Faculty of Science, Macquarie University, Australia
Brian.jersky@mq.edu.au

In more and more cases in Australia and worldwide, several subjects that fall under the general area of Statistics are taught in high school classes, and sometimes even earlier. It is thus of interest to test what first year University students know about the subject on entry to an Introductory Statistics class. Of perhaps even more interest is to test such students after an introductory statistics class.
This paper presents a case study in which the author pre- and post-tested a class of first year students taking an Introductory Statistics class taught in the (northern) Spring semester at a small, private, liberal arts college in the San Francisco Bay area of California.

Because of the inherent difficulties involved in testing students in a “low stakes” testing environment, where the marks obtained do not “count” towards a final grade, the students were told that the tests would enable the instructor to more usefully tailor instruction to their needs in the class for the pre-test, and would help their peers in future classes for the post-test.

The results of the tests were interesting in several ways, though none of them alone would perhaps be particularly surprising to an experienced instructor. This paper analyses the pre- and post-tests and comments on the results obtained.

The instrument used for the test was constructed using the testing resources obtained from the ARTIST website. This website was constructed based on as NSF grant to develop useful and reliable instruments for introductory statistics topics and concepts, and the paper briefly describes these and how they were used in the case study.
Essentially, what was found was that initial misperceptions about several topics were entrenched before the class even began, and remain so afterwards in many cases. Perhaps this points to areas where instructors should focus more attention, and also to areas where less attention might be beneficial.

UNDERSTANDING OF SAMPLING VARIABILITY: CONFRONTING STUDENTS' MISCONCEPTION IN RELATION TO SAMPLING VARIABILITY

JAZAYERI Mitra, LAI Jerry, FIDLER Fiona and CUMMING Geoff

Department of Mathematics and Statistics, School of Psychological Science
La Trobe University
m.jazayeri@latrobe.edu.au

Psychology students typically have difficulty understanding the relationship between the sample size and variability. For example, they often do not understand that variability of sample means decreases as the size of the samples drawn increases and instead believe variability increases with sample size. Understanding this is crucial to their understanding of standard error. This paper studies the effects of applying conceptual change theory to a class of psychology students studying statistics, focussing on understanding the relationship between sample size and variability. The methodology implemented to evaluate the effects of applying conceptual change theory to psychology students comprised of two phases. The first assessed the students understanding at the beginning of the semester, to establish a baseline for comparison in subsequent analysis. In the second phase, the students were divided into two groups; standard and intervention; to evaluate the changes in their level of understanding and the overall success of the intervention teaching strategy. In order to ascertain the misconception of sampling variation, we designed a teaching intervention which directly confronted the variability misconception of the students. The intervention was developed by applying cognitive conflict as the basis for conceptual change. We tested this intervention on a class of n=185, 'Statistics for Psychology' students. Approximately half the students received the direct confrontation and the other half received 'standard training'. The students completed a questionnaire on the topic, before the intervention, immediately after and at the end of semester (approx 6 weeks follow-up). The result of the Pearson Chi-Square test of independence indicates moderate evidence regarding the association between the level of the understanding of the students about the variability of the sample means for different sample sizes and the method of teaching.

AUTHOR INDEX