



ROYAL
STATISTICAL
SOCIETY

DATA | EVIDENCE | DECISIONS



Abstracts booklet

RSS International Conference • 7 – 10 September 2015 • Exeter University
www.rss.org.uk/conference2015

Abstracts are ordered in session order for oral presentations followed by poster presentations

1.1 Contributed – Medical – Randomised Trials

Tuesday 8 September 9.30am – 10.30am

Sample size calculation for cluster randomised trials with ordinal outcomes

Clare Rutterford, Andrew Copas, Sandra Eldridge
Queen Mary University London

Background

A common approach to sample size calculation for cluster randomised trials is to calculate the sample size assuming individual randomisation and multiply this by the design effect. Calculation of the design effect requires knowledge of the cluster size and intraclass correlation coefficient (ICC), a measure of the extent of clustering. It is not yet clear how well the design effect method works for ordinal outcomes.

Objectives

To evaluate the performance of the design effect approach in sample size calculation for cluster randomised trials with ordinal outcomes and to provide recommendations for the calculation of the ICC.

Methods

The performance of the design effect method was evaluated across a range of scenarios, chosen to reflect the characteristics of trials using ordinal outcomes.

For each scenario sample size was calculated using the design effect approach with three alternative estimators for the ICC: an ANOVA based estimate, a Kappa type estimate and the ICC on the assumed underlying continuous outcome. A thousand datasets of appropriate size were generated and each analysed via a random-effects model. Empirical power was calculated as the proportion of datasets with significant result.

The simulation studies were extended to explore performance under alternative analysis methods and variable cluster sizes.

Conclusions

The use of the design effect works well with ordinal data, except in the case of a small number of clusters. Power calculations using the ANOVA estimate of the ICC performed adequately, and that ICC is simple to calculate.

1.1 Contributed – Medical – Randomised Trials

Tuesday 8 September 9.30am – 10.30am

Sample size calculation for Stepped Wedge Trials

*Rumana Omar, Andrew Copas, Gareth Ambler, Emma Beard
University College London*

Stepped Wedge Trial (SWT) a variant of a clustered randomised trial (CRT), embed additional complications with regards to design and analysis which should be addressed adequately in its sample size calculations to ensure valid conclusions about the intervention effect. We performed a review of analytical methods for sample size calculations for SWT and how sample size has been calculated for SWTs in practice. We discuss the assumptions underlying currently used methods and their validity. We propose simulation based methods which offer more flexibility by relaxing assumptions as well as incorporating the features of the SWT design in the sample size calculation more appropriately. We performed a simulation study to compare simulation-based approaches with analytical methods and the efficiencies of the SWT and standard CRT designs for continuous and binary outcomes with cross-sectional and closed cohort SWT designs.

Sample sizes obtained using the simulation approach and analytical methods were similar when the underlying assumptions of the methods were aligned. However, the simulation-based approaches are more useful in handling the specific features of SWT such as the closed cohort design, random intervention effect and time effect. The sample size for SWT was found to be relatively insensitive to intracluster correlations (ICC) values. Failure to account for a potential time effect in the sample size calculation reduced the power of the study considerably. In line with existing literature, the SWT was found to require a smaller number of clusters/sample size compared to CRT when the ICC is relatively high (e.g. greater than 0.1).

1.1 Contributed – Medical – Randomised Trials

Tuesday 8 September 9.30am – 10.30am

Avoiding Type 0 Errors

*Patricia Gunning, Alberto Alvarez-Iglesias, John Newell
NUI Galway*

A time consuming part of statistical analysis is building analysis datasets for use in statistical procedures. Whether it is deriving new variables, restructuring datasets or combining data from multiple data sources, the work to create an analysis data set can be complex. Raw data are rarely, if ever, fit for analysis in a ready rectangular shaped state. For example, data to be analysed are often stored in a database system often logically organised from the point of view of data management or IT but not necessarily from the view point of statistical analysis.

Collaborating statisticians should be able to conceptualise and programme appropriate transformations of raw data creating valid datasets for analysis based on complex, multi-sourced linked datasets. Of all the steps in the statistical inference process, the performance of this important initial step is usually not given the attention it deserves. Errors at this stage may render results invalid. Catching this pre-analysis Type 0 error is critical.

This presentation will take a historical tour of data transformation techniques used to create analysis datasets from brute-force approaches to recently developed dedicated packages in R, an open-source statistical software. The dplyr and other packages will be illustrated using novel examples showing the powerful and elegant functionality offered by R.

References:

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>.

Wickham, H. and Francois, R. (2014). dplyr: dplyr: a grammar of data manipulation. R package version 0.2. <http://CRAN.R-project.org/package=dplyr>.

1.2 Contributed – Official – Administrative Data

Tuesday 8 September 9.30am – 10.30am

The assurance of administrative data - a proportionate approach

Emily Gleeson, Penny Babb
UK Statistics Authority

Administrative data collections provide an attractive means of obtaining generally cheap and comprehensive data for producing official statistics. But they can present substantial challenges, coming from systems with unique definitions and needs. Statistical producers may have little involvement in the data collection.

The Statistics Authority has developed a proportionate approach to the quality assurance of administrative data - with three levels of assurance: basic, enhanced and comprehensive. It centres on the level of risk associated with the data and degree of public good represented by the statistics, to decide which level of assurance is required.

The QA toolkit helps statistical producers identify appropriate areas of practice to assure the data for their particular statistics, and will be used in the Assessment of official statistics by the Authority. The four practice areas demonstrate the need for the quality assurance of statistics obtained from administrative data to extend beyond the checks made by statistical producers on the data they receive. They should also have a good understanding how the data are compiled and why, have established effective communication links with data supply partners, and understand their partners' data quality processes and standards.

Essential to this good statistical practice is the publication of clear guidance for users, setting out the strengths and limitations of the administrative data, to support the further use of the official statistics.

This presentation will describe the toolkit, which has been refined following a consultation with statistical producers, and provide examples of its use.

1.2 Contributed – Official – Administrative Data

Tuesday 8 September 9.30am – 10.30am

Delivering early benefits and trial outputs using administrative data

Pete Jones

Office for National Statistics, UK

Following the Government's endorsement of the National Statistician's recommendation on 'The census and future provision of population statistics in England and Wales', the ONS Beyond 2011 Programme has been closed and replaced by the new Census Transformation Programme.

The new programme is focusing on developing the strategies and plans needed for delivery of the following major strands of work:-

1. an online census in 2021;
2. integrated statistical outputs that make use of administrative data and surveys in conjunction with census;
3. a recommendation for the future provision of population statistics beyond 2021.

Strand 3 continues with research carried out in the Beyond 2011 Programme exploring the potential of administrative data and surveys as a future alternative to traditional Census taking beyond 2021. Building upon the concept of 'Statistical Population Datasets' derived through anonymous linkage of multiple administrative sources, the ONS plans to release a series of annual 'trial output' statistics to deliver early benefits and engage users with the development and evaluation of methods.

'Trial outputs' are intended to illustrate what might be realised from administrative data, in particular the range and frequency of outputs, and the potential for small area statistics. The first release will focus on local authority population counts at age/sex level. Subsequent annual releases will aspire to produce smaller area population counts and additional outputs on household estimates, income and ethnicity, subject to data access and quality. This presentation will outline ONS plans to deliver trial outputs in the run up to the 2021 Census.

1.2 Contributed – Official – Administrative Data

Tuesday 8 September 9.30am – 10.30am

Assessing uncertainty when combining administrative data to estimate population counts

Dilek Yildiz, Peter W. F. Smith
University of Southampton

The aim of the research is to develop a methodology to make use of already collected administrative data to estimate population counts for England and Wales in the absence of a traditional census. The key goals are to produce population counts by age, sex and region (local authority) by combining administrative data, and estimate associated measures of precision. Previous research (Yildiz and Smith 2015) showed that combining administrative data by log linear models with offsets in a classical framework is useful to decrease, to a certain extent, the bias in administrative data sources. However, they did not provide the estimates of uncertainty for the final estimates. Therefore, this research moves away from the frequentist approach, extends the previous work, and focuses on combining data sources in a probabilistic framework. To illustrate the proposed methodology the aggregate level patient register is combined with auxiliary data by log linear models with offsets in a Bayesian framework.

1.3 Contributed – Environment – Extremes

Tuesday 8 September 9.30am – 10.30am

How extreme can storms get in space and time?

Theodoros Economou, David Stephenson, Christopher Ferro
University of Exeter

Extreme North Atlantic extra-tropical cyclones are a major source of risk for society. These natural hazards cause much damage and insurance loss in Europe due to extreme wind speeds and flooding. An important scientific question therefore is: how extreme can these cyclones become? Unlike hurricanes, there are no simple thermodynamic arguments for the minimum pressure achievable by extra-tropical cyclones. In the absence of reliable physical arguments, we use a statistical model to provide an answer to this question. The challenges in implementing such a model include: 1) Extremal storm properties vary spatio-temporally, 2) events occur irregularly in space and 3) events are rare in many regions.

A Bayesian hierarchical extreme value model is used to characterise storm pressure subceedances below the local 0.1 quantile, using data on extra-tropical cyclones over the Atlantic and Europe extracted from 6-hourly re-analyses data in 1979–2009. Spatial variation in the extremal properties of the cyclones is captured using a 150 cell spatial regularisation, latitude as a covariate, and spatial random effects. The North Atlantic Oscillation (NAO) is also used as a covariate and is found to have a significant effect on intensifying extremal storm behaviour, especially over Northern Europe and the Iberian peninsula. The model is implemented using MCMC and extends existing spatial extreme value methodology by including temporal covariates with spatially varying coefficients. Estimates of lower bounds on storm pressure are typically 10–50 hPa below the minimum values observed for historical storms with largest differences occurring when the NAO index is positive.

1.3 Contributed – Environment – Extremes

Tuesday 8 September 9.30am – 10.30am

Bayesian inference for non-stationary extremes

David Randell, Kathryn Turnbull, Philip Jonathan Shell

Specification of realistic environmental design conditions for marine structures is of fundamental importance to their reliability over time. Design conditions for extreme waves and storm severity are typically estimated by extreme value analysis of time series of measured or hindcast significant wave height, H_s . This analysis is complicated by two effects. Firstly, H_s exhibits temporal dependence. Secondly, the characteristics of storm peak H_s are non-stationary with respect to multiple covariates, particularly wave direction and season.

We develop directional-seasonal return values for storm peak significant wave height for a location in the South China Sea by estimation of, and simulation under a non-stationary extreme value model. Extreme value model parameters vary with direction and season, and are parameterised in terms of (tensor products of) B-spline bases for direction and season. Inference is carried out using MCMC, exploiting the Manifold Metropolis Adjusted Langevin Algorithm (mMALA) for improved convergence and mixing of chains. Return value estimates from Bayesian inference are compared with those from maximum penalised likelihood estimation.

1.3 Contributed – Environment – Extremes

Tuesday 8 September 9.30am – 10.30am

ABC model selection for spatial extremes

Tony Pettitt, Chris Drovandi, Xing Lee
Queensland University

The work here presents an approximate Bayesian computation (ABC) based method for model selection between spatial extreme models. ABC algorithms rely on evaluation of discrepancies between model simulations and the observed data rather than explicit evaluations of computationally expensive or intractable likelihood functions, as is the case for spatial extreme models.

The method was developed and applied to annual maximal temperature data from 26 weather stations dispersed around South Australia. Agriculture and horticulture are strong industries in South Australia which depend heavily on the particular climate conditions of the state. In particular, South Australia experiences dry and hot weather which generally support these industries. However, extreme weather events can be detrimental to these industries. For example, bushfires, a consequence of extremely high temperatures coupled with low humidity, have been reported on numerous occasions in South Australia. It is thus important to be able to confidently estimate climate conditions particularly in areas further away from weather stations in order to better manage and safeguard such industries.

The results from the ABC model selection showed a clear preference for the Brown-Resnick model out of the six spatial extreme models considered. The fitted model exhibited strong correlation between spatial locations in South Australia. In addition, the model parameters were also sufficiently well-estimated given the strong preference for the model, circumventing the need to re-estimate the model parameters separately.

1.4 Contributed - Social Science – Ageing & Inequality

Tuesday 8 September 9.30am – 10.30am

Cognitive ageing in Great Britain in the new century: Can the recent cohorts postpone it?

Gindo Tampubolon

University of Manchester

The ageing populations in high income countries are facing a challenge to the cognitive functioning of older people. Dementia in these countries is set to be the third major burden of disease even as the period of older age continues to extend with the increase in life expectancy. This challenge demands an understanding of how cognition changes over an extended period. But studying cognitive ageing faces a difficulty arising from the fact that older respondents are liable to drop out before study completion. Three hypotheses are tested: trajectories of cognitive ageing show improvement, not only decline, over a decade; also the trajectories are lifted by secular increases in cognition across cohorts; and lastly the trajectories are susceptible to distortion due to drop out.

Using the English Longitudinal Study of Ageing 2002-2013, this paper draws trajectories of cognitive scores and episodic memory of Britons aged 50 and over. It applied growth curve models and joint models to estimate trajectories of cognition. The analysis found that levels of cognitive scores and of episodic memory follow a curvilinear shape in later life, contradicting a steady decline beyond the age of 50. The findings also revealed secular improvement in cognitive ageing such that as a cohort is being replaced cognitive scores levels improve.

In conclusion this study conveys an optimistic message that today older age is not necessarily a period dominated by cognitive decline. In identifying behavioural factors associated with better cognitive ageing, it raises possibilities of mustering adequate response to the cognition challenge.

1.4 Contributed - Social Science – Ageing & Inequality

Tuesday 8 September 9.30am – 10.30am

The dynamical approach to inequality

Alexander Adamou, Ole Peters
London Mathematical Laboratory

The non-ergodic nature of multiplicative growth manifests itself in a difference between time and ensemble average growth rates. The time average growth rate of wealth (or income) reflects the typical economic experience of an individual, whereas the ensemble average growth rate reflects the economic experience of the aggregated population. Statistics about the latter are widely available in the form of GDP growth figures: these inform policy-making and are frequently reported in the media. Data on the former would provide a more complete picture of economic growth but are lacking. The difference between the two growth rates measures changes in economic inequality in a society. This insight suggests a new and fundamental definition of inequality, which we present here as part of a general framework based on underlying wealth dynamics. In the context of multiplicative growth, the dynamical approach reveals Theil's (1967) second measure of inequality - the mean logarithmic deviation - as the natural measure of inequality.

1.5 Contributed – Methods for Handling Data Dependencies

Tuesday 8 September 9.30am – 10.30am

Singular working covariance models for informatively observed longitudinal data

Daniel Farewell, Chao Huang
Cardiff University

We consider the analysis of longitudinal data where the number, and timing, of observations may be informative about their recorded values. We argue that suitably defined 'marginal' longitudinal parameters can be of substantive interest if an arbitrary measurement schedule could (in principle) be imposed upon a subject. Within the framework of generalised estimating equations, we demonstrate that solutions obtained using singular, low rank, working covariance structures are consistent for these marginal parameters. Crucially, only structural assumptions are required; in particular, there is no need for parametric models relating the chance of obtaining a given observation pattern to its corresponding longitudinal responses. There remains considerable flexibility in the choice of path to the singular working covariance, and efficiency is gained by choosing this path using the observed data. Our approach compares favourably to existing alternatives, including inverse probability weighting and joint modelling. We illustrate these new methods by analysing longitudinal data from a clinical trial that investigated novel forms of peritoneal dialysis.

1.5 Contributed – Methods for Handling Data Dependencies

Tuesday 8 September 9.30am – 10.30am

Flexible copula sample selection models for count responses

Karol Wyszynski, Giampiero Marra
University College London

Non-random sample selection arises when observations do not come from a random sample. Instead, individuals select themselves into (or out of) the sample on the basis of observed and unobserved characteristics. In this case, standard statistical models such as linear or logistic regression will yield biased and inconsistent estimates. This problem can be addressed using sample selection models. For instance, McGovern et al. (2015) find that the HIV prevalence for men in Zambia amounts to 12% when using standard models and 20% when sample selection models are used in the analysis. In the methodological literature a lot of attention has been given to sample selection models with continuous response. At the same time, not much work has been attributed to sample selection models with count response. Meanwhile, many studies from the field of economics, sociology, finance and medicine can be affected by the presence of non-random sample selection. We will consider a copula-based sample selection model for count data with flexible covariate effects. Theoretical and numerical issues will be discussed. Hence, model comparison methods and inferential procedures will be presented. The model will be illustrated on hospitalization data originating from the United States Veterans Administration.

1.5 Contributed – Methods for Handling Data Dependencies

Tuesday 8 September 9.30am – 10.30am

Frequency domain inference for rapidly evolving nonstationary processes

Arthur P. Guillaumin, Sofia C. Olhede, Adam M. Sykulski, Jeffrey J. Early, Jonathan M. Lilly
University College London

A time series is a set of observations tracked over time, often described by their correlation structure. A traditional assumption in time series analysis is that the time series is stationary, or exhibits homogeneous correlation in time. In real-world applications this is often unrealistic. We therefore need to develop methods for understanding evolving processes.

In this talk we extend concepts from the theory of stationary processes to a certain class of nonstationary processes, which vary sufficiently fast in time to make regular theory for nonstationary processes not applicable. This is illustrated with a time-varying autoregressive process.

For stationary processes, inference using maximum likelihood in the time domain is not always computationally feasible. A well-established and fast approximation, known as the Whittle likelihood, is computed in the frequency domain. A benefit of such a frequency-domain approach is the ability to describe oscillatory patterns. We discuss how the Whittle likelihood must and can be extended to our given class of nonstationary processes.

We show the relevance of our method with an application to surface velocities of ocean currents, and focus on the Coriolis effect, responsible for a time-varying oscillatory pattern. The nonstationarity of the Coriolis effect is particularly pronounced in the Gulf stream, where velocities change rapidly. Understanding the Gulf Stream is important to global climate modelling as changes to this phenomenon have drastic effects on NorthWest Europe. In addition the Global Drifter Programme is a massive collection of time-indexed oceanographic observations, necessitating a methodology which balances statistical optimality and computational tractability.

1.6 Contributed – School performance & work placements

Tuesday 8 September 9.30am – 10.30am

Does School Performance Add Up?

Paul Askew

Huntingdonshire District Council

Annual spending on education is £90bn¹, 8% of total national spending, reflecting the potential impact on personal life chances and general economic prosperity. There are detailed measurement systems in place in schools (at “key stages”) to help measure educational effectiveness, becoming national and official statistics. Those measurement systems provide a very large volume of detailed data, and there is a range of quite different reporting and visualisation solutions intended to help make sense of that data at local level.

The quality of the school level understanding of that performance - by managers, teachers, parents, school governors – is a critical factor that enables the right local interventions. Because of the potential volume of data and the range of potential visualisations, a strong and consistent understanding of the issues and successes can be difficult to achieve.

This paper details an Academy Trust solution that refocused the use of performance data to be more strongly purpose and framework driven. This has enabled: (1) a more predictive approach to allow earlier interventions; (2) stronger integration between pupil, class, teacher, year group, and school performance to differentiate both pupil and systemic issues and successes; (3) standard and intuitive (but sufficiently flexible) data visualisation that has enabled consistency and familiarity; (4) a longitudinal and multi-dimensional data tracking of the pupil journey through their school career providing a stronger pupil and cohort centric approach (5) stronger engagement with stakeholders generating a more focussed understanding of performance.

¹ HM Treasury, Public Expenditure Statistical Analysis 2014. Table 5.2)

1.6 Contributed – School performance & work placements

Tuesday 8 September 9.30am – 10.30am

How statistics determine examination results in England

Tom Benton

Cambridge Assessment

This talk will provide a detailed description of how information on pupils' achievement in primary school examinations at age eleven is currently used to determine what percentage of them achieve the top grades in GCSE examinations in each subject at age 16. The talk will discuss the strengths and weaknesses of this approach based upon a recent evaluation of the technique for the qualifications regulator Ofqual. On the one hand, using resampling methods, it is possible to show that the standard errors associated with the recommended pass rates are kept fairly low. However, by comparing the recommended pass rates from the current technique to slightly improved approaches, based on more powerful data (that is, concurrent attainment rather than prior attainment) this talk will also show how the technique may fail to correctly adjust for changes in national attainment at primary school and also underestimate the extent to which pass rates should differ between different exam boards offering alternative versions of a GCSE in a given subject. Finally the talk will discuss the reasons why, despite its weaknesses, the regulator prefers to give more weight to statistical approaches to determining examination pass rates and less weight to evidence from expert judgement.

1.6 Contributed – School performance & work placements

Tuesday 8 September 9.30am – 10.30am

Do work placements work?: How the University of Manchester Q-Step Centre is making data analysis real for undergraduates by putting them into the workplace

Jackie Carter

University of Manchester

The University of Manchester has one of the 15 nationally funded Q-Step Centres. We have two broad aims; to get more social science students using numbers in their undergraduate studies, and to train the quantitative social researchers of the future.

In support of these aims we have introduced a programme of work placements in the summer of the second year. Successful students undertake a real world, data-driven, social research 'project' over an 8-week period. In Summer 2014 we placed 19 students in Manchester and London in think tanks, polling organisations, market research companies, consultancy and university research groups and university policy and planning departments. The focus of these placements was on skills development. All students benefited with some returning to take up a quantitative dissertation in their 3rd year, and some now entering quantitative professions or applying for statistics and social research methods postgraduate courses. In summer 2015 we will place 46 students in similar organisations, expanding to banks and the third sector and covering more cities.

This presentation will cover an evaluation of the first year of the Q-Step work placements, analysing in-depth interviews with all 19 students and their summer employers, to highlight emergent themes from both perspectives. We explore questions such as: Are we teaching social science students what employers need? How do work placements develop statistical skills? How can lessons learned feed back into the data analysis curriculum for social sciences? How much maths do students need to do data analysis? Do work placements work?

2.1 Invited – Statistical causality & applications in medical research: opportunities and challenges

Tuesday 8 September 12 noon – 1.20pm

Causal inference in epidemiology: the challenge of strong confounding

Rhian Daniel, Stijn Vansteelandt
London School of Hygiene & Tropical Medicine

This talk focuses mainly on settings with a single exposure, single outcome and a number of potential confounders.

Typically, in such settings, causal inference methods focus on a causal estimand that compares two hypothetical worlds: one in which every individual is exposed and one in which every individual is unexposed. This is also the comparison implicitly made by traditional approaches to analysing epidemiological data if the estimated parameters are interpreted causally. We discuss the challenges of drawing inferences about such an ambitious comparison in the presence of strong confounding, ie when some individuals, on the basis of their confounders are either very likely or very unlikely to be exposed. In particular, we discuss the unjustified extrapolation on which inferences rely in this setting when they are based on models for the outcome conditional on exposure and confounders.

We advocate some alternative causal estimands along with an alternative estimation methods based on modelling the exposure given confounders, which together insulate against extrapolation bias. We illustrate these ideas using examples from cardiovascular epidemiology.

2.1 Invited – Statistical causality & applications in medical research: opportunities and challenges

Tuesday 8 September 12 noon – 1.20pm

Adding value to large RCTs: applications of causal inference methods

Deborah Ford, Angela Crook, Sarah Walker
University College London

Applications of causal inference methods to RCTs have generally focused on adjusting for non-compliance to the randomised interventions. In contrast, we present three applications which address important non-randomised questions. The DART trial randomised HIV-infected African adults initiating anti-HIV treatment to CD4 versus no CD4 monitoring. Twelve-weekly CD4s were performed in both groups, but only returned for management in one group. Cotrimoxazole (antibiotic) prophylaxis was prescribed at the treating clinician's discretion in both groups. We used marginal structural models with inverse-probability treatment weights to estimate causal effects of cotrimoxazole on outcomes. Current use of cotrimoxazole significantly reduced mortality. During the trial, switch to second-line anti-HIV treatment was recommended for low CD4 (if returned) or WHO stage-4 clinical events. In an application of dynamic marginal structural models across pooled randomised groups we showed that 24-weekly CD4-monitoring with switch for low CD4 resulted in similar survival to 12-weekly monitoring and that a single CD4 after 48 weeks first-line treatment (to identify non-responding patients) improved survival over clinical monitoring. The MDP-301 trial randomised African women to a candidate microbicide gel to prevent HIV-acquisition versus placebo. We estimated the effects of hormonal contraceptive use on HIV-acquisition by artificially censoring women who changed contraception or became pregnant after baseline and applying joint censoring weights to account for biases this may have introduced. Injectable depo-medroxyprogesterone increased HIV risk. RCTs offer large amounts of diverse high quality data which may be successfully exploited to answer questions about patient management beyond the randomised question.

2.1 Invited – Statistical causality & applications in medical research: opportunities and challenges

Tuesday 8 September 12 noon – 1.20pm

Scope of applications of causal inference methods in RCTs: A systematic review of the literature.

*Daphne Kounali, Deborah Ford, Sarah Walker, Margaret May
University of Bristol*

Purpose: To systematically review published studies of RCT data to identify applications exploiting causal inference methods to deal with time-dependent confounding which addressed questions that were not answered by randomisation.

Methods: Systematic review of published papers identified in Medline, Premedline, Embase, Cochrane Library, and Web of Science from 1986 onwards employing a forward citation search of five seminal papers and keyword search including the following terms: time-varying confounding; causal inference; marginal structural models; inverse probability weighting; g-estimation; g-formula; structural nested models. Studies were excluded if inverse probability weighting was used: to balance baseline characteristics; to adjust for loss to follow-up; to adjust for non-compliance to randomised treatment. Studies where the exposure could not be assigned were also excluded.

Results: From the 2773 studies initially retrieved only 19 studies were relevant. The majority (12/19) estimated the causal effect of concomitant medication on outcome; whilst the remainder were concerned with post-randomization treatment regimens (sequential-treatments) including comparison of second-line treatments dependent on response to first-line treatment, effects of timing of second-line treatment and treatment dosing. Examples were found in cardiovascular disease (7), HIV (5), cancer (5) and mental health (2). Only twelve studies were reported in clinical journals.

Examples of studies which exploit causal methodology using RCT data remain limited, despite the growth in methodological development. Practical difficulties may include lack of power to estimate non-randomised effects and limited variability in patient management within an RCT. Researchers should consider exploring additional clinical questions within RCTs to maximise the resources expended on running them.

2.2 Invited – The use of admin data in the National Energy Efficiency Data framework (NEED)

Tuesday 8 September 12 noon – 1.20pm

What can DECC's NEED data tell us about levels and change in domestic electricity consumption?

Ben Anderson
University of Southampton

There is broad agreement that dwelling characteristics, environmental and climatic factors, the 'need' for cleanliness and comfort, together with 'what people do' within dwellings drive UK domestic gas consumption. However much less is known about the factors affecting electricity consumption, especially where heat is provided by other sources of energy. This paper will present preliminary analysis of the 2014 End User License File (EULF) sample of the domestic NEED data obtained through the UK Data Service. It will start by presenting analysis of year on year variation in electricity consumption at the dwelling level and will then focus on the extent to which floor area (as a proxy for the number of occupants) and other dwelling characteristics, deprivation levels, main heat fuel and energy efficiency ratings provide robust predictors of electricity consumption. Analysis of other national survey sources will then be used to understand the kinds of households we might expect to find in these dwelling types and thus to form currently untestable hypotheses to explain the results found in NEED. The paper will conclude with an outline of current proposals being taken forward through the Administrative Data Research Centre England (ADRC-E) to link the domestic NEED data to a range of other sources to provide robust information on dwelling occupants and, to the extent possible, their energy consuming habits in order to test these hypotheses.

2.2 Invited – The use of admin data in the National Energy Efficiency Data framework (NEED)

Tuesday 8 September 12 noon – 1.20pm

The National Energy Efficiency Data-framework: its creation and insights

Gregory Haigh

Department of Energy and Climate Change, UK

The National Energy Efficiency Data-framework (NEED) won the 2014 RSS Statistical Excellence in Official Statistics award for its innovative approach to linking diverse data, providing detailed insights into energy saving potential and allowing householders to assess their energy use.

NEED draws together household-level administrative data from a number of different sources, including: metered energy consumption; property and household attributes; and details of energy efficiency measures installed. Combining this data enables valuable insights to be made, for example the real-world impact that cavity wall insulation installed across the housing stock has had on energy consumption, while controlling for other potential factors.

This presentation will:

- explain how NEED was created, using a custom address matching algorithm to allow the disparate datasets to be linked
- outline some of the interesting results the analysis has revealed
- explain how these insights have been used in Government for monitoring and evaluation and policy development

2.2 Invited – The use of admin data in the National Energy Efficiency Data framework (NEED)

Tuesday 8 September 12 noon – 1.20pm

Building a Model of Domestic Gas Consumption Using Administrative Data

Tamas Borbely

Department of Energy and Climate Change, UK

The predictors of domestic energy use are of great theoretical and practical interest to a variety of groups, from academics to diverse industries and policy-makers. The socio-demographic, psychological, and property-related characteristics that drive households' electricity and gas consumption have a rich research literature, but the available evidence pertaining to the relative importance and interactions of these characteristics is equivocal. The Department of Energy and Climate Change (DECC) is in a unique position to contribute to this debate as it has access to administrative data from a wide range of sources that are linked up through the National Energy Efficiency Data Framework (NEED). NEED contains actual electricity and gas meter readings for over 26 million residential properties, along with detailed property information and modelled household characteristics for a subset of over four million records. In this presentation, we describe the research project conducted by DECC aimed at using NEED to develop a model of domestic gas use. We outline methodological challenges, including the reconciliation of data obtained from different sources and dealing with the complex interactions between predictor variables. Various statistical techniques were evaluated for suitability in the project; these are also briefly described. Results are discussed in terms of their relevance to the academic debate around patterns in domestic energy use as well as to the development and evaluation of energy policy.

2.3 Invited – Environmental Extremes

Tuesday 8 September 12 noon – 1.20pm

Modelling extremes on river networks

*Peiman Asadi, Anthony Davison, Sebastian Engelke
EPFL, Switzerland*

Max-stable processes are the natural extension of the classical extreme-value distributions to the functional setting, and they are increasingly widely used to estimate probabilities of complex extreme events. In this talk I shall discuss how to broaden them from the usual setting in which dependence varies according to functions of Euclidean distance to the situation in which extreme river discharges at two locations on a river network may be dependent, either because the locations are flow-connected or because of common meteorological events. In the former case dependence depends on river distance, and in the second it depends on the hydrological distance between the locations, either of which may be very different from their Euclidean distance. Inference for the model parameters is performed using a multivariate threshold likelihood, which is shown by simulation to work well. The ideas are illustrated with data from the upper Danube basin.

2.3 Invited – Environmental Extremes

Tuesday 8 September 12 noon – 1.20pm

Extreme value methods for heatwave risk assessment

Jonathan Tawn, Hugo Winter
Lancaster University

Heatwaves are a type of extreme event defined as a period of persistent hot temperatures. They can lead to substantial increases in death rates and costs to the farming industry. Examples of two such major events are the focus of this talk, namely the 2003 European heatwave (40,000 deaths and 13M Euro loss) and the 2009 south-eastern Australia heatwave (350 deaths and record temperatures).

Many previous extreme value studies have focused on modelling the maximum intensity of an event without considering the effect that the duration and spatial extent might have. From a modelling perspective this represents a focus on the marginal distribution while ignoring the extremal dependence structure. In this talk a conditional approach for multivariate extremes will be presented. This provides a framework to model the spatial and temporal dependence and for incorporating covariates. These models are used to assess how the risks of the two major heatwaves listed above differ with climate change (a 1°C warming in the global temperature) and with natural climate phenomena (El Nino Southern Oscillation).

2.3 Invited – Environmental Extremes

Tuesday 8 September 12 noon – 1.20pm

Temporal downscaling for rainfall extremes: models for Depth Duration Frequency curves

David Walshaw
Newcastle University

We consider a problem motivated by a serious practical question: what do the observed extremes of daily aggregate rainfall tell us about the extremal behaviour of hourly rainfall? Very often only daily rainfall records are available at a particular location, but return levels for hourly values are needed for design safety.

The return level curves for a function integrated over different time windows, considered together, form what are known as Depth Duration Frequency curves (DDFs). In this talk we review the fairly limited literature on this topic. We go on to consider the options for modelling the joint behaviour of extremes aggregated over different time windows at a single location, starting with exploratory analysis in the form of downscaling (regression) models to relate the extreme value parameters/return levels at individual locations.

We consider the opportunities presented by a new and extensive hourly data set of UK rainfall for improved estimation of DDFs at individual locations, through spatial models.

2.4 Invited – Training the quantitative social scientists of tomorrow

Tuesday 8 September 12 noon – 1.20pm

Statistics or applied data science? From chi-squared to (geo)computation via some pretty maps

Richard Harris
University of Bristol

There has been widespread lament about the decline of quantitative methods teaching and learning within UK social science. The consequence is said to be harmful to research, to the employability of students and to citizenship: if people cannot understand the data they are presented with, they cannot debate it. The ESRC, HEFCE, British Academy and Nuffield Foundation have invested considerable money in resolving this situation, with the most visible projects currently the national Q-Step Centres and the Secondary Data Analysis Initiative. Within schools, maths is being strengthened, with the possibility of core maths or A/AS level being compulsory for all pupils. In Universities, the review of benchmark statements provides opportunity to strengthen the role of quantitative methods within curricula. Yet, there is no clear understanding of what actually we mean by quantitative methods - is it pure maths, applied data handling, 19th/20th century statistics or basic numeracy? Is it a return to the past or does quantitative methods mean something different today that has more to do with data science, geocomputation and visual communication than it does with more traditional statistics? Drawing on work that has been undertaken in geography, I will discuss some of the ambiguities, the vogue for embedding quantitative methods within curricula, why this is not a return to positivism, and what actually a modern syllabus in quantitative methods might look like.

2.4 Invited – Training the quantitative social scientists of tomorrow

Tuesday 8 September 12 noon – 1.20pm

Igniting the statistical spark in the social sciences

Rhys Jones
Cardiff University

A major strand of the Cardiff Q step centres work focuses on developing new courses in Social Analytics in the FE and schools sectors across Wales. Social Analytics is the scientific investigation of social processes, using statistical techniques as an exploratory lens. Developments include: delivering a pilot scheme in Social Analytics for FE/high school students, the creation of new Level-3 qualifications (A' level equivalent) in Social Analytics with the Welsh Joint Education Committee and Agored Cymru, and the expansion of a teacher/FE lecturer network. This talk will explore a novel and integrative approach of engaging with the FE/schools sectors to develop new national courses in Social Analytics, with consideration given to the potential impacts on the higher education sector.

2.4 Invited – Training the quantitative social scientists of tomorrow

Tuesday 8 September 12 noon – 1.20pm

Telling Stories with Data: Integrating Data Analysis into the Social Science Curriculum

Susan Banducci, Travis Coan
University of Exeter

Improving the data analysis skills of UK undergraduates is now a major priority in UK social sciences, with £19m invested in the QStep programme. In this session I present on the approach taken to quantitative methods teaching at the Exeter QStep Centre. I showcase the teaching strategies, tools and techniques that we are developing in three UG programmes in Sociology, Criminology and Political Science. We start from the position that quantitative social science data and the skills to manage and analyse them are a core part of the social sciences. An undergraduate social science degree that does not introduce students to these approaches and skills is too narrow. At Exeter Q-Step we have built our UG curriculum on the principles of embedding data analysis skills into substantive modules focusing on the following elements: telling stories with data, data visualisation, and causal inference.

2.5 Invited – Recent advances in computational statistics using Holonomic Gradient Methods

Tuesday 8 September 12 noon – 1.20pm

Holonomic gradient methods for the Fisher-Bingham normalising constants

*Alfred Kume, Tomonari Sei
University of Tokyo*

Holonomic function theory is successfully implemented in a series of recent papers to efficiently calculate the normalizing constant and perform likelihood estimation for the Fisher-Bingham distributions. A key ingredient for establishing the standard holonomic gradient algorithms is the calculation of the Pfaffian equations. So far the present papers either calculate these symbolically or apply certain methods to simplify this process. In this talk, we show the explicit form of the Pfaffian equations using the expressions from Laplace inversion methods. This aids the implementation of the holonomic algorithms for these problems and enables their adjustments for the degenerate cases. They include important sub-classes of the Fisher-Bingham distributions such as Bingham and Kent families. As a result, an exact and more efficient algorithm for likelihood inference for these distributions is generated. We illustrate efficiency of the method with real data sets.

2.5 Invited – Recent advances in computational statistics using Holonomic Gradient Methods

Tuesday 8 September 12 noon – 1.20pm

Origin and developments of holonomic gradient method

Akimichi Takemura
University of Tokyo

I will present origin and developments of holonomic gradient method (HGM). HGM utilizes differential equations for a class of functions called "holonomic functions". Many standard probability density functions, such as the normal density, are holonomic and this method can be used. Since its inception, it has been used for many examples and found to work very well. Derivation of the differential equation can be automated by the use of Grobner basis theory for the ring of differential operators.

2.5 Invited – Recent advances in computational statistics using Holonomic Gradient Methods

Tuesday 8 September 12 noon – 1.20pm

An Estimation Procedure Based on the Information Geometry of Dually Flat Spaces

Yoshihiro Hirose
University of Tokyo

We show a geometrical method for parameter estimation. Our method is motivated by the least angle regression (LARS) algorithm. Our bisector regression (BR) algorithm is applicable to the exponential family of distributions (including the generalized linear models) while the LARS algorithm is available only in the normal linear regression. In the BR algorithm, the bisector of an angle plays an important role, which is a similar situation to that of the LARS algorithm. The BR estimator moves along an extension of the bisector of an angle in the model manifold. The BR algorithm outputs the sequence of estimates. Each estimate belongs to each submodel of a nested family of submodels. The length of the sequence is equal to the number of the parameters to be estimated. This fact means that the BR method narrows down the candidate submodels drastically while the number of all candidates is much larger than that of the parameters.

In the BR algorithm, the holonomic gradient method is available in updating the BR estimator. The trajectory of the BR estimator is characterized by a system of ordinary differential equations. The holonomic gradient method is expected to help the BR algorithm work well even when the BR algorithm does not work by itself.

This is a joint work with Prof. Komaki and Prof. Sei at The University of Tokyo.

3.1 Invited – Perspectives on Stratified Medicine

Tuesday 8 September 2.10pm – 3.30pm

An introduction to stratified medicine

Dan Jackson, Brian Tom
MRC Biostatistics Unit, Cambridge

In this brief introduction to the session, we will explain what is meant by "stratified medicine". We will outline some main areas of current research in this area and will discuss some of the key literature. This introduction will be followed by three further, and much more in-depth, talks about stratified medicine.

3.1 Invited – Perspectives on Stratified Medicine

Tuesday 8 September 2.10pm – 3.30pm

High-dimensional statistical approaches for heterogeneous molecular data in cancer medicine

Nicolas Städler, Frank Dondelinger, Steven Hill, Rehan Akbani, Gordon Mills, Sach Mukherjee
MRC Biostatistics Unit, Cambridge

Molecular interplay plays a central role in basic and disease biology. Patterns of interplay are thought to differ between biological contexts, such as cell type, tissue type, or disease state. Many stratified high-throughput studies now span multiple such contexts and the data may therefore also be stratified with respect to patterns of interplay. This motivates a need for statistical approaches that can cope with molecular data with possibly unknown stratification, which are heterogeneous in a multivariate sense.

In this work, we exploit recent advances in high-dimensional statistics to put forward tools for analysing heterogeneous molecular data in cancer medicine. We model the data using Gaussian graphical models, and develop two useful techniques based on estimation of partial correlations using the graphical lasso: a two-sample test that captures differences in molecular interplay or networks, and a mixture model clustering approach that simultaneously learns a sensible stratification and multivariate network models that are specific to each stratum.

We demonstrate the characteristics of our methods using an in-depth simulation study, and proceed to apply them to proteomic data from The Cancer Genome Atlas (TCGA) "pan-cancer" study. Our analysis of the TCGA data provides formal statistical evidence that protein networks differ significantly by cancer type. Furthermore, we show how multivariate models can be used to refine cancer subtypes and learn associated networks.

Our results demonstrate the challenges involved in truly multivariate analysis of stratified molecular data and the substantive gains that high-dimensional methods can offer in this setting.

3.1 Invited – Perspectives on Stratified Medicine

Tuesday 8 September 2.10pm – 3.30pm

Pharmacogenetics

John Whittaker

GSK

Genetic variants provide plausible stratifying variables for both the efficacy and safety of medicines, and much study has produced some important and interesting examples. In this talk I review briefly the current state of knowledge and show several case studies. Based on these I argue that we should expect genotyping of clinical trials to discover relatively few clinically actionable genetic stratifiers, and that a more promising approach may be to use genetic/genomic data to suggest indications early in the clinical development process. Finally, I discuss briefly other potential 'omic stratifiers of medicines. Throughout I will avoid technical detail, both of genetics and of statistics.

3.1 Invited – Perspectives on Stratified Medicine

Tuesday 8 September 2.10pm – 3.30pm

Statistical Regulation in Stratified Medicine

Deborah Ashby, Rob Hemmings
Imperial College London

Medicines have been regulated in the UK and other jurisdictions for over fifty years. As the science underpinning new medicines has evolved, so has the regulation. ICH-E9, has endured as the main guidance for statistical principles, despite being written twenty years ago, although there are now many more detailed guidelines on particular topics.

In this talk, we review the challenges that stratified medicine presents from a statistical perspective, with recent examples. We explore the extent to which these challenges are in fact effectively covered by existing guidance, and highlight areas where fresh thinking in design, analysis or interpretation may be required by those making submissions and by regulators.

3.3 Invited – Free Data

Tuesday 8 September 2.10pm – 3.30pm

Balancing data security and utility - a perspective from the Office for National Statistics as both a user and owner of Government data

Lucy Vickers

Office for National Statistics, UK

The Office for National Statistics owns and collects statistical data, for example the Census and Labour Force Survey. We are committed to opening up this data for research whenever legally able to do so, and use various techniques and policies to protect confidentiality to achieve this safely. Increasingly, we also use administrative data sources from across Government for research and statistical production. We have developed sophisticated data security infrastructure, methods, expertise and procedures to facilitate our own access to these data, especially where the data are particularly sensitive.

In addition, we are collaborating with the University of Southampton and Bloomsbury group to run the Admin Data Research Centre for England (ADRC-E). The Admin Data Research Network, of which the centre is a part, is intended to facilitate researcher access to linked Government owned admin data. To achieve this, the network has developed a data security model whereby only trusted third party data linkers (ONS perform this role for ADRC-E) access dataset identifiers to determine how two datasets link together. Researchers only access linked attributes that have been disclosure checked, in a controlled secure environment, after linkage is complete and the identifiers have been removed.

We will give a short presentation explaining some of these areas further, and use our expertise in to contribute to the panel's discussion on balancing data security and utility.

3.3 Invited – Free Data

Tuesday 8 September 2.10pm – 3.30pm

Free data: Turning numbers into information

Emily Cadman
Financial Times

The amount of data in the public sphere is growing exponentially. Putting aside the ONS' catalogue, data.gov.uk estimates it has 20,000 data sets on its site, barely a month goes by without some public institution launching a new open data initiative, and developers are always finding new datasets they can construct.

But despite the volume of resources available there is a paucity of actual information.

What journalists are looking for from data is insight: what have we learnt about the world that we did not know before? Has a policy or an event had an impact (positive or negative) we can now see in hard figures? Can that impact be quantified in terms of geography, gender, race or class?

With so much data available, and the most interesting information typically requiring multiple data sets to be combined, the key is to start with the right type of question.

Emily Cadman is the FT's UK economics reporter. Data, and the stories you can tell with it, are at the heart of her professional life. She has worked in journalism for over a decade in a variety of roles, ranging from head of the FT's interactive data visualisation team to web editor on a variety of trade magazine titles.

3.3 Invited – Free Data

Tuesday 8 September 2.10pm – 3.30pm

Sharing individual participant data from clinical trials: Why, How, and Can we do better?

Catrin Tudur Smith
University of Liverpool

The availability of Individual Participant Data (IPD) from clinical trials can facilitate comprehensive, robust analyses beyond those planned in the original protocol. This can lead to new clinical insights, help to inform the design of future clinical trials and help to advance the development of new methodology. However, there are challenges for clinical trialists attempting to share IPD as well as for researchers trying to access this data.

In this talk I will draw upon personal experiences of accessing and using IPD and will discuss the advantages, alternative models for sharing IPD, how the landscape of clinical trial data sharing is changing, and how more could be done to facilitate the responsible sharing of IPD.

3.3 Invited – Free Data

Tuesday 8 September 2.10pm – 3.30pm

Using the Secure Data Service to Access Administrative Data from the UK

Catherine Robinson, Chiara Rosazza Bondibene, Geoff Mason
University of Kent

Administrative data for the UK has been routinely accessed via a special license, whereby access is granted to anonymised individuals' records under restricted conditions but largely based on trust. Usually, data are provided directly to the researcher in anonymised form. However, in the case of businesses, anonymity is harder to achieve, and the data are potentially commercially sensitive. The UK has a number of competing models for secure access. This presentation will focus on the practical use of the Secure Data Service (SDS). This is ESRC-supported access to data produced by the Office for National Statistics and other government departments (such as BIS and BERR), administered by the UK Data Service at the University of Essex. As well as providing an overview of the type of data currently stored, the advantages and limitations of data accessed through the SDS are considered here, using a recent research project as an example of what can be undertaken. We argue that this system enables researchers in remote parts of the UK (and potentially outside) access to data that would otherwise be restricted to those working in London and other major cities. Moreover, the nature of the environment supports collaborative research. Overall, the access arrangements for the SDS help to support diversity within the research environment on a number of levels.

3.4 Invited – Cutting Edge Statistical Modelling in the Social Sciences

Tuesday 8 September 2.10pm – 3.30pm

A multilevel longitudinal analysis of rich-poor achievement gaps in English schools

George Leckie
University of Bristol

Education systems across the world are increasingly focusing on narrowing ethnic and other achievement gaps seen in their country's schools. We focus on England where since 2010 the Government have held schools publicly accountable for their rich-poor achievement gaps (socioeconomically disadvantaged pupils vs. their more advantaged peers) via publication of these statistics in national school league tables. It is hoped that this will also lead to the identification of successful schools and therefore policies which can be replicated broadly.

We present a multilevel longitudinal analysis of schools' achievement gaps to illustrate severe limitations in the Government's publication of these statistics. Specifically, we fit multilevel binary logistic regression models to student exam success where we include random school coefficients on student disadvantaged status, exam year, and their interaction. We fit all models using Markov chain Monte Carlo methods as implemented in the MLwiN software.

The results show a slight reduction in the national rich-poor achievement gap, but substantial variation across schools. Unfortunately, many instances of apparent success in narrowing the gap are driven by a lowering of the performance of advantage pupils rather than any particular improvement of disadvantaged pupils. Other instances are driven more by the changing prior attainments of the two pupil groups rather than any genuine improvement in schools' policies and practices. More fundamentally, the small numbers of students involved lead inferences about individual schools to be severely unreliable.

We conclude that the published statistics are of limited value for their stated school accountability and school improvement purposes.

3.4 Invited – Cutting Edge Statistical Modelling in the Social Sciences

Tuesday 8 September 2.10pm – 3.30pm

Allowing for Measurement errors in generalised linear models.

Harvey Goldstein
University of Bristol

In 2003 an analysis by Leon Feinstein had important effects on educational policy. He claimed to show that there was a very large increasing discrepancy between children from low and high SES groups as they got older. His analysis, however, ignored a large measurement error variance associated with the test scores used and adjusting for this substantially modifies his inferences. The talk introduces a model that allows such measurement errors to be incorporated into the model specification so that the underlying (true) relationship of interest can be estimated.

3.4 Invited – Cutting Edge Statistical Modelling in the Social Sciences

Tuesday 8 September 2.10pm – 3.30pm

Understanding English Higher Education: Making modelling accessible

Mark Gittoes

Higher Education Funding Council for England

A number of projects where hierarchical modelling has been used to better understand the patterns and trends seen in English higher education will be discussed. The presentation will also highlight different approaches that have been used to make the results of this modelling accessible to non-technical audiences. Topics that will be covered include the regional variation seen in higher education participation, and differential performance and outcomes of different groups in degree studies.

3.5 Invited – Forecast Evaluation

Tuesday 8 September 2.10pm – 3.30pm

Evaluation of prognostic value in medical research

Thomas Gerds

University of Copenhagen

The many new biomarkers that abound in cardiology and other branches of medicine have given rise to the novel-marker-question. It is a familiar observation that the incremental information offered by a new marker is very often disappointingly small when it is added to an already well discriminating battery of clinical data. For example, quite large odds ratios are required to considerably improve the area under the ROC. The AUC and other traditional performance statistics have therefore been accused of being insensitive to new information (an undeserved accusation, in our view).

3.5 Invited – Forecast Evaluation

Tuesday 8 September 2.10pm – 3.30pm

Assessing the calibration of high-dimensional ensemble forecasts using rank histograms

*Thordis Thorarinsdottir, Michael Scheuerer, Christopher Heinz
Norwegian Computing Centre*

Any decision making process that relies on a probabilistic forecast of future events necessarily requires a calibrated forecast. We propose new methods for empirically assessing forecast calibration in a multivariate setting where the probabilistic forecast is given by an ensemble of equally probable forecast scenarios. Multivariate properties are mapped to a single dimension through a pre-rank function and the calibration is subsequently assessed visually through a histogram of the ranks of the observation's pre-ranks. Average ranking assigns a pre-rank based on the average univariate rank while band depth ranking employs the concept of functional band depth where the centrality of the observation within the forecast ensemble is assessed. Several simulation examples and a case study of temperature forecast trajectories demonstrate that both multivariate ranking methods can successfully detect various sources of miscalibration and scale efficiently to high dimensional settings.

3.6 Invited – The use of technology and web resources in teaching statistics

Tuesday 8 September 2.10pm – 3.30pm

Web resources and computer based tools for teaching statistics

*Gillian Lancaster, Svetlana Tishkovskaya, Andrew Titman
Lancaster University*

Internet technologies and computer based tools have opened up new possibilities for the teaching of statistics. Even though the use of online learning environments in education is expanding, there are very few studies that have systematically assessed the teaching methods and resources available to teachers of applied statistics at the tertiary level. This talk will overview the current state of the art in web-based learning environments in statistical education, rating and categorising resources into helpful groupings, to help teachers find and implement useful technologies in a variety of innovative ways for the enhancement of statistics teaching. In addition the use of computer based tools such as personal response systems (PRS) or 'clickers' will be highlighted with an illustration. We use PRS at MSc level in intermittent places within the Principles of Epidemiology module to check students' understanding and knowledge of the topic using multiple choice questions within each lecture. Module design and student feedback will be discussed.

3.6 Invited – The use of technology and web resources in teaching statistics

Tuesday 8 September 2.10pm – 3.30pm

Use of technology in a distance learning statistics module

Tom Fanshawe
University of Oxford

Distance learning modules in statistics are becoming increasingly popular as a means of attracting students from a global audience. They have particular value in providing continuing professional development, when students may not have the time or resources to attend face-to-face, lecture-style teaching sessions. This talk will review the technology available to support distance-learning modules, and highlight their advantages and disadvantages. Topics discussed will include the provision of bespoke online learning material (both text-based and using multimedia) and efficient use of technology to improve interaction between students and tutors, for example through discussion forums and chat rooms. Using such content requires the tutor to be aware that students may have different preferences for the format in which course materials are presented, and therefore offering a variety of media is recommended. Experiences of successful approaches will be discussed using as an example an introductory statistics module in a Masters course in Evidence-Based Health Care. Distance learning teaching styles will be compared with those adopted in face-to-face courses and those that use a combination of face-to-face and distance learning.

3.6 Invited – The use of technology and web resources in teaching statistics

Tuesday 8 September 2.10pm – 3.30pm

Interactive Statistics with The Open University

Rachel Hilliam, Carol Calvert
Open University

The Open University has over forty years' experience in distance learning and has transformed how its material is presented to students with advancing technology. Although television programs fronted by men in kipper ties at 5am in the morning are a thing of the past, new ways of engaging large numbers of students with the learning requirements of our courses have been developed whilst the core material, mainly in printed form, retains the excellent standard for which The Open University is famous. In 2013 a new introductory statistics course was presented to students from a wide range of disciplines and backgrounds for the first time. This talk will highlight the wide variety of interactive resources used by students on the course which have demonstrably enhanced the traditional print based teaching associated with statistics. These include over 70 screencasts demonstrating key teaching examples, interactive computer animations, activities using the software package (Minitab), a database of practice questions containing randomly generated variants to allow multiple attempts at problems and forums for cohorts of students and tutors to interact. All this is embedded in the OU's virtual learning environment, which enable student activity to be tracked so that appropriate interventions can be targeted to students who need extra support. Over 2000 students have passed the module so far and student feedback is very positive.

3.6 Invited – The use of technology and web resources in teaching statistics

Tuesday 8 September 2.10pm – 3.30pm

Data to Insight: MOOCing around with statistics education

Chris Wild

University of Auckland

"Data to Insight" is an introduction to statistical data analysis MOOC (online open course, but not so massive) taught on the UK's FutureLearn platform late last year and due to run again late this year. Course completers ranged from members of a small high-school physics-honours class to PhD researchers from many areas, and from journalists, linguists and arts administrators to economists, data managers, marketers and scientists. It was pleasing to see a large contingent of high-school mathematics teachers, less so to see so many "data analysts". Completers were extremely enthusiastic about the course.

In addition to being a MOOC introducing its students to statistical data analysis, "Data to Insight" prototyped a much-further-much-faster, more-data-more-quickly Introductory Statistics Course. The most novel acceleration strategies used were: being intensely visual and driving all argument off things you can see supplemented by metaphor; building software solutions (including powerful online visualisation and analysis software) that prevent "how do I get this out of the software?" limiting the speed at which students can encounter new situations and new ideas; and finding some powerful, conceptually-undemanding "extender-capabilities" that immediately open much wider horizons.

We will speak about the course, the software, the MOOC environment and educational lessons learned from working in a MOOC environment for both online and classroom teaching. We will highlight lessons learned from making the course videos with a team of professionals, from the use of online quizzes and from online communication and discussion with and between students.

4.1 Contributed – Medical – Clinical Trials

Tuesday 8 September 3.50pm – 4.50pm

On the Generalizability of Experimental Results

Chaitra Nagaraja, Dylan Small
Fordham University

While assignment to treatment and control groups are randomized in experiments (e.g., clinical trials), selection into the trial itself is most often not. This makes it more difficult to generalize experimental results to the wider population. Currently, this issue is handled by estimating the probability that any given person in the population decides to participate in the trial using methods such as propensity scores. These techniques require the use of auxiliary, population-level data such as gender, age, and race distributions. However, such models leave open the (likely) possibility that key variables were omitted from the analysis or modeled incorrectly.

Our novel approach is to use an instrumental variable, such as distance from the trial site, to more accurately model these trial participation probabilities. Such a variable avoids the omission bias by being correlated with trial participation but not directly with trial outcomes or other covariates. This unique feature results in an improved method to determine whether experimental results can indeed be generalized.

4.1 Contributed – Medical – Clinical Trials

Tuesday 8 September 3.50pm – 4.50pm

Reproducible Research in the Analysis of Regulated Trials

Alberto Alvarez-Iglesias, Patricia Gunning, John Newell
NUI Galway

The efficacy and safety of new therapies/drugs are generally evaluated with the conduct of regulated trials. Clinical trials are heavily regulated to protect the rights, safety and well-being of subjects and to assure the credibility of the results reported. To comply with regulations, all study documents, data and procedures must be traceable and reported findings should be reproducible. Although there are numerous guidelines on the content, evaluation and reporting of statistical findings for these trials, there is a lack of practical guidelines on how to implement these principles.

In this presentation, we will discuss reproducible research in the analysis of clinical trials and will present a customised in-house state-of-the-art process that we built. The system is a reproducible, transparent, validated and robust reporting system that is in compliance with current regulations. It incorporates statistical best practice starting from the time data are received to the generation of the statistical report. This process is implemented using Sweave, a tool in R which embeds code and text in a single file, allowing for easy re-generation of reports when the data or analyses change. Version control is incorporated; changes to files are recorded over time allowing specific versions to be recalled later if needed. Results are verified by an independent statistician using commercially validated software.

References

Peng, R.D. (2009). Reproducible research and Biostatistics. *Biostatistics*, 10, 3, pp. 405-408.
Gandrud, C. (2014). *Reproducible Research with R and RStudio*. Boca Raton FL: CRC Press.

4.1 Contributed – Medical – Clinical Trials

Tuesday 8 September 3.50pm – 4.50pm

Modelling the impact of uncertainty in the study endpoint for a clinical trial: application to prostate screening.

Stephen Walter
McMaster University

The European Randomised Study of Prostate Cancer Screening (ERSPC) has the objective of estimating the impact of PSA-based screening on prostate cancer mortality, and currently it shows an approximately 21% reduction in this endpoint. In contrast, a similar trial (PLCO) in the United States shows no reduction in prostate cancer mortality. Accurate adjudication of causes of death is challenging, and so we analysed the European data to assess if unreliability in the adjudication of causes of death might have influenced the results.

Because there is no "gold standard" for the cause of death, several latent class models (LCM's) were developed to determine: 1) the accuracy of individual adjudicators; 2) if their accuracy varied; and 3) if adjudication error rates were differential between the randomised study arms. Finally we used the model results to evaluate if correcting for variability in adjudication might affect the ERSPC conclusions.

There was some variation between adjudicators, but the pairwise agreement between them was generally very good, and there was only limited evidence of asymmetry in their classification of causes of death. There was no consistent evidence of differential accuracy by trial arm. Using the LCM results, we generated alternative estimates of the study odds ratios, corrected for observer unreliability, but they were quite similar to the original estimates.

We conclude that bias arising from adjudication inaccuracy was unlikely. This analysis represents an interesting example of a useful approach to assessing potential bias in trial results when determination of the main endpoint is uncertain.

4.2 Contributed – Official – Quality & Progress

Tuesday 8 September 3.50pm – 4.50pm

Reviewing official inflation statistics

*Paul Johnson, Eric Crane, Rosie Foster
Office for National Statistics, UK*

In July 2013, Paul Johnson, Director of the Institute for Fiscal Studies, was asked to lead a review of the range of consumer price statistics produced by the Office for National Statistics (ONS). These include the Retail Prices Index (RPI) and Consumer Prices Index (CPI). They are used, amongst other things, to uprate pensions and benefits, in the setting of interest rates, and in determining the interest payable on about a quarter of the UK government's debt.

The review was supported by a small team that worked alongside the Division in ONS responsible for consumer price statistics. The review worked with ONS, external experts and stakeholders to understand the key issues around these statistics. The review also commissioned several pieces of research. The review's final report was published in January 2015, and is set to shape the production of consumer price statistics for years to come.

This session will offer an insight into a review of one of the highest profile sets of statistics in the UK. It will also briefly cover the findings of the review.

4.2 Contributed – Official – Quality & Progress

Tuesday 8 September 3.50pm – 4.50pm

Implementation of Quality Dimensions within the Office for National Statistics

Tess Carter

Office for National Statistics, UK

This presentation will outline the dimensions of quality before providing emphasis on the ways in which the Office for National Statistics (ONS) has improved outputs in accordance with each dimension.

Quality is made up of several dimensions which are Relevance, Accuracy, Timeliness/punctuality, Accessibility/clarity and coherence/comparability. These will be exemplified using various examples from the outputs produced by various branches within ONS.

Some of the examples discussed within the presentation will include:

- Relevance- User event incorporated several different types of output including short stories, interactive maps, an article and a listicle to provide a package of work that catered to all stakeholders.
- Accuracy- Foreign Direct Investment team has updated methodology to incorporate changes from BPM5 to BPM6 and increase sample size.
- Timeliness/punctuality- Annual Business Survey team conducted an imputation review.
- Accessibility/clarity- Outwards Foreign AffiliaTe Survey team published using new HDML format.
- Coherence/comparability- Research and Development team have improved explanation methods using infographics.

The presentation will highlight the ways in which ONS implements the quality dimensions and will provide inspiration for others to do the same.

4.2 Contributed – Official – Quality & Progress

Tuesday 8 September 3.50pm – 4.50pm

Nigeria Statistical System: The Evolution, Progress and Challenges

Obafemi Keshinro, Olusanya Olubusoye
University of Lagos

Official statistics is a key player in every national transformation agenda. This study examined the evolution of the Nigerian Statistical System (NSS) from 1928 to date. The historical development was categorized into four stages: Phase I-Early Stage (Pre-1957 Statistics Act), Phase II: Development State (1957 - 1988), Phase III: Reorganization and Consolidation (1988-2007) and Phase IV: where the National Strategy for the Development of Statistics (NSDS) evolved (2007 to date). The paper investigated and discusses some of the emerging challenges facing the National statistical System. Findings showed an astronomical increase in demand for data between 2005 and 2012 with the main drivers for the demand of data being the result of weakening growth and the search for new opportunities. The increased demand for accountability from citizens is another major reason for the increased production of data. However, progress remains insufficient, suppliers/users are not well educated, funding/support remains dependent on sources outside the country, the process of programming and annual monitoring and evaluation remains incomplete and does not include all producers, confidence in the NSS is low and governance remains weak. The status of implementation of the NSDS was found to be limited by existing ICT infrastructure. This study should enable the orientation of policy formulations targeted towards the development of an improved NSS.

Keywords: Statistics, Transformation, Development, Strategies, Policy Formulations

4.3 Contributed – Environment – Natural Hazards

Tuesday 8 September 3.50pm – 4.50pm

Statistical simulation of natural hazard event footprints.

Ben Youngman
University of Exeter

Natural hazard events, including hurricanes, floods and earthquakes, are often represented as footprints, which give a spatial representation of events, sometimes aggregated over their duration. Tools to simulate footprints can form part of catastrophe models, which are used by reinsurers for loss estimation. Hazard footprint simulation models are often built similarly to -- or based on -- regional climate or weather forecasting models, and so can have heavy computational requirements. Furthermore, such models can poorly capture the extremity of hazards, or their frequency, leading to unreliable loss estimates.

This talk presents a statistical model for simulating natural hazard footprints that combines methods from extreme value theory and geostatistics; the former gives accurate and theoretically sound representation of an event's extremity, while the latter captures spatial dependence, which governs an event's spatial extent. Due to its relatively small computational cost, our model can quickly simulate thousands of years worth of footprints at high resolution using just a desktop computer. We study European windstorms, which are important events to understand: for example, in 2014 at US\$2.2bn windstorm Ela caused the greatest insured loss of all natural catastrophes occurring in Europe. We achieve realistic simulations of extreme European windstorm footprints, which are formally validated in terms of their extremal and spatial properties. This approach to footprint simulation is intended to offer a new perspective on catastrophe modelling.

4.3 Contributed – Environment – Natural Hazards

Tuesday 8 September 3.50pm – 4.50pm

A decision theoretic approach for issuing natural hazard warnings

*David Stephenson, Theo Economou, Rob Neale, Ken Mylne
University of Exeter*

Warnings for natural hazards help improve societal resilience and are a good example of decision-making under uncertainty. Despite been proposed more than 20 years ago for flood warnings (Krzysztofowicz, 1993), no operational warning systems for ANY natural hazards currently appear to make use of well-established Bayesian decision theory.

In this pilot study, we propose a simple framework for hazard warning systems based on Bayesian decision theory. The framework uses probabilistic predictions of the state of nature in conjunction with a loss function to issue optimal warnings with respect to expected loss. The loss function is constructed as the sum of the losses of three hypothetical agents: the forecaster, the protector who pays for preventative actions, and the insurer who pays for ensuing losses after each event. Statistical methods are also proposed and compared for evaluating the probability of the state of nature given real-time forecast information.

The approach is illustrated using 24 hour ahead 4-state warnings of severe daily precipitation totals over the UK and is compared to the current heuristic severe weather warning system used operationally by the UK Met Office for warnings in the 2012/13 and 2013/14 winters. Weather warnings based on the proposed decision theoretic approach were found to be different than those issued by the Met Office, and resulted in substantially smaller losses. A Generalised Additive Model used to recalibrate ensemble weather forecasts was shown to provide improved probability predictions and thereby further reduce losses.

4.3 Contributed – Environment – Natural Hazards

Tuesday 8 September 3.50pm – 4.50pm

A bivariate copula model for extremal windstorm footprints

Laura Dawkins, David Stephenson, Julia Roberts, Christopher Ferro, Ken Mylne, Ben Youngman
University of Exeter

Windstorms are one of the most destructive natural hazards in Europe. Each windstorm has an associated footprint, defined as the maximum 3 second wind-gust to occur at each location in the domain in the 72 hour period covering the passage of the storm. The most damaging historical windstorms have had footprints with differing characteristics, some have a large area of low wind-gust speeds while others have a smaller area of higher wind-gust speeds.

A statistical model for the windstorm footprint is being developed to investigate which of these characteristics is most important for determining the loss a windstorm will cause. The initial step is to develop a bivariate model, which must have the potential to be extended to model the entire footprint, and must be able to represent realistic joint losses.

The bivariate distribution of damaging wind-gusts at two locations can be represented by the marginal distributions of wind-gusts and their mutual dependence, explained by a copula. Extreme value theory is used to fit the marginal distributions and an appropriate copula model is selected after investigating the extremal dependence between wind-gusts at pairs of locations. Comparing the empirical joint probability of loss with those calculated using the model validates the ability for the model to represent realistic joint losses.

The left-truncated Generalised Extreme Value distribution is used for the margins and the Gaussian copula explains the dependence in the model, which is found to successfully represent realistic joint losses, and is relatively simple to extend using Geostatistics.

4.4 Contributed – Social Science – Terrorism & Risk Management

Tuesday 8 September 3.50pm – 4.50pm

Learning to save lives after terrorist attacks and industrial accidents - allowing for individual variation in performance

Neil Spencer

University of Hertfordshire

Paramedics who need to save lives after a terrorist attack or industrial accident may have to perform intubation (inserting a plastic tube into the windpipe) whilst wearing a CBRN (Chemical, Biological, Radiological, Nuclear) protective suit. When learning how to undertake this task, completion times are initially high but gradually reduce as the paramedic gains experience and, after a change point, they flatten off.

From a statistical point of view, each individual's set of times can be modelled as a broken stick/hockey stick/linear spline regression. As is common with longitudinal studies, multilevel modelling can be used to allow individuals' performances to vary over time in different ways. With broken stick regression, it lets individuals have different initial abilities, different levels of improvement and different final abilities.

However, much previous work has restricted individuals to having a change point at one of a small number of discrete occasions, often imposing one "best fit" change point on all individuals. Alternatives that use Bayesian hierarchical methods may still restrict the number of possible change points, may be sensitive to choices of priors or may not be implementable with readily available software.

This paper presents a method for allowing individual change points that does not suffer from these difficulties. Treating the change point as a parameter which can vary between individuals and using a parameterisation which leaves the model in a readily interpretable form, it uses a Bayesian hierarchical model with noninformative priors that can run in software such as OpenBUGS.

4.4 Contributed – Social Science – Terrorism & Risk Management

Tuesday 8 September 3.50pm – 4.50pm

Statistical modelling of a terrorist network

Murray Aitkin, Duy Vu, Brian Francis
University of Melbourne

The recent book "Disrupting Dark Networks (Everion 2012) gives an extensive analysis of the Noordin Top terrorist network. The identification of subgroups of actors in the network is ambiguous, with different algorithms giving different numbers and compositions of subgroups.

This talk investigates the group structure through the latent class model and a new Bayesian model comparison method for the number of latent classes. Under one version of the model it clearly identifies a group containing the leaders and organisers, and the group structure suggests a hierarchy of leaders, trainers and "footsoldiers" who carry out the attacks.

Simulations show that the new method can correctly identify the number of latent classes, given sufficient data.

4.4 Contributed – Social Science – Terrorism & Risk Management

Tuesday 8 September 3.50pm – 4.50pm

Scoring Method of Enterprise Risk Management

*Jake Ansell, Juthamon Sithipolvanichgul
University of Edinburgh*

Enterprise Risk Management (ERM) is seen as an holistic approach to ensure a good risk management strategy for companies to help minimise potential pitfalls and improve long term business sustainability. However questions still arise whether ERM implementation impacts on a firm's performance. Past studies have shown no consensus that ERM does increase firm performance as advocated by regulators and business advisors. So the issue exists as to whether ERM implementation has been adequately assessed. An alternative measurement of ERM implementation is proposed. The measurement is based on standardised integrative scoring. The relationship between the proposed measurement and firm performance is then considered taking account of appropriate control variables. Using data from the Thailand Stock Exchange it was found implementing ERM can improve firm performance in term of Tobin's Q, ROE and ROA.

4.5 Invited – Prize Winners YSM 2015 Conference

Tuesday 8 September 3.50pm – 4.50pm

Understanding patient-level variation in antibiotic use within primary care

Nick Beckley

University College London

Antibiotics have been one of the greatest breakthroughs in modern medicine, transforming bacterial infections from life-threatening to routinely treatable. Although bacteria evolve to become resistant to these antibiotics over time, this shouldn't be a major health threat if administered responsibly. However, bacterial resistance has been hugely accelerated in recent years due to inappropriate use, and is considered to be one of the biggest global scientific issues today.

Work to date has focused primarily on what might influence antibiotic prescription rates from the general practice perspective, with resultant "blanket" recommendations uniformly applicable to all patients. In contrast, little is understood about patient-level factors that may influence antibiotic prescription rates. This work aims to increase our understanding of the patient characteristics that may influence higher prescription rates.

Antibiotic prescription records were extracted from The Health Improvement Network (THIN) primary care databases from 2011-2014. The cumulative number of prescriptions for each patient were stratified by age, sex and social deprivation. Curves were fitted to these data, and values for the top 10%, 5%, 1% and 0.1% were interpolated as measures of "high usage". Bootstrap confidence intervals were also generated to compare the levels of "high usage" between groups.

Our results showed little variation between sex and social deprivation classes, but high levels of variation between age classes, with the very young and older patients more likely to be higher users of antibiotics. These results can inform GPs and help to reduce the number of antibiotic prescriptions in high users.

4.5 Invited – Prize Winners YSM 2015 Conference

Tuesday 8 September 3.50pm – 4.50pm

Defining Rural England

Matthew Dray

Department for Environment, Food and Rural Affairs

Rural areas cover 85 per cent of England's land surface and are home to over 9 million people. Opportunities, challenges and barriers can differ between rural and urban settlements, affecting businesses, access to services and quality of life. Rural and urban areas should therefore be differentiated when analysing social and economic statistics. So how can the government define rurality? This presentation will discuss how Defra's Rural Statistics team and its collaborators classified and mapped England's rural and urban areas. The definition of 'rural' depends on spatial scale. At its most detailed, the Rural-Urban Classification assigns small-scale geographic units to one of six rural or four urban categories. This information can be used to inform classifications at larger spatial scales, such as the local authority level. The classifications are now available as open data products via the Office for National Statistics and have recently been deployed to help distribute European funding.

4.5 Invited – Prize Winners YSM 2015 Conference

Tuesday 8 September 3.50pm – 4.50pm

Approaches to calibration in survey sampling.

*Gareth Davies, Jonathan Gillard, Anatoly Zhigljavsky
Cardiff University*

The technique of calibrating sample surveys seeks to assign optimal weights to members of a sample in order to form good estimates of population statistics. This talk will outline various calibration methods adopted by statistical offices throughout the world. The calibration procedure seeks to find new weights that are close as possible to the initial design weights, subject to various constraints. The constraints may lead to negative or extreme weights. Several approaches to overcome these issues will be motivated, including the use of range restrictions and relaxing the constraints. The talk will conclude with an application to estimating the UK unemployment rate from the Labour Force Survey.

4.6 Contributed – Teaching Statistics – Tools & engagement

Tuesday 8 September 3.50pm – 4.50pm

Developing Interactive eBooks and an Analysis Assistant to Teach and Apply Modern Quantitative Methods

*Richard Parker, Danus Michaelides, Christopher Charlton, Luc Moreau, William Browne
University Of Bristol*

Whilst a large range of valuable training resources are available to those interested in learning quantitative techniques, few discuss how statistics are conducted in practice by working analysts. Advances in technology, however, have opened up the possibility of using more interactive tools to develop such resources. We have conducted interviews with quantitative researchers from a variety of disciplines, and are collaborating with them to produce interactive eBooks based on case studies they have chosen. These are written using our Stat-JR package which can interoperate with a variety of other statistical software, and can thus allow users to explore how a given analysis might be conducted in a range of different packages. The resulting eBooks will form a library of case studies that those newer to the field can use as a learning tool. Furthermore, the insights and experiences of our participating analysts, together with that of the project's investigators, will also feed into the design of a statistical analysis assistant, which will integrate the eBook system with software designed to run analysis workflows.

4.6 Contributed – Teaching Statistics – Tools & engagement

Tuesday 8 September 3.50pm – 4.50pm

Data visualization as a tool for engagement with statistics and research

Daniel Ellis, Tim Hobbs
Dartington Social Research Unit

Data visualization is not a new phenomenon, but is perhaps an under utilized tool. There is a huge amount of data available to us from across a wide range of disciplines. How we access, use and interact with the data made available is often an after thought to the publication of it. Interactive visualisations of research data can broaden your audience as well as helping your existing audience engage with, and understand it, better. Communicating data and research findings well should be as important as conducting good quality analysis and research.

The Dartington Social Research Unit has recently designed and disseminated several data visualizations with the aim of better engaging commissioners and policy makers in the wealth of complex child well-being data collected through our epidemiological work.

Different datasets require different types of design to best communicate the key messages trapped within them and there is no catchall template that always works. There is however a relatively simple process that can be used to guide the development of a successful data visualization. This always starts with the data and works up through key messages the data may hold, how best to represent this in graphical form, and finally how to display and present this to the audience who would be using the data in question.

We would like to share our experience of making data more widely accessible, concluding with the benefits and pitfalls of doing so, namely the difficulty of data retaining its original context.

4.6 Contributed – Teaching Statistics – Tools & engagement

Tuesday 8 September 3.50pm – 4.50pm

Teaching with national and cross-national survey data resources with the UK Data Service: access, analysis and visualisation using online tools and resources

Hersh Mann
University of Essex

The UK Data service provides online access to over 6000 data collections based on censuses, social surveys, international databanks and other sources. Using online tools such as Nesstar and UKDS.Stat, researchers are able to search for, locate and browse through millions of variables and visualise statistical information within their web browsers without the need for specialist software. Teachers of quantitative methods within the social sciences often report reluctance amongst some students to engage with numerical information because they lack confidence in their own statistical literacy. The UK Data Service has responded to provide a suite of online pages focused on teaching along with a series of supporting resources designed to provide a gentle introduction to quantitative social research. These include downloadable guides, workbooks, video tutorials and scaled down 'teaching datasets' based on well-known large scale surveys. Our feedback from teachers shows that students are more engaged with these resources because they are easily accessible and, because they are often based on national data sources which have a large media impact, provide a key stimulus for class-based discussion and critical appraisal. The teaching case studies we have published online demonstrate the value of providing specialist teaching material which allays the fears of anxious students and provides them with a foundation upon which they can test hypotheses and become confident users of survey data and official statistics

4.7 Invited - EPSRC Statistics & Applied Probability review and its follow up

Tuesday 8 September 3.50pm – 4.50pm

Chris White
EPSRC

The International Review of Mathematical Sciences (IRM) in 2010 highlighted that UK Statistics research was in a weakened condition, despite the many areas of excellence, and that measures needed to be taken in the coming years to avoid the loss of the UK's international stature. EPSRC's strategy to shape support for Statistics and Applied Probability research area over the past few years was in part due to the recommendations made by the IRM panel. Statistics and Applied Probability is one of 15 'grow areas' across the whole EPSRC remit, and the only grow area in the Mathematical Sciences portfolio. The EPSRC Mathematical Sciences theme last year selected Statistics and Applied Probability as an area to review in detail. Whilst the area has grown since 2011, the funding environment is changing and EPSRC wanted to consult the community on how things have changed over the past few years and identify potential actions for the future. A Review Day was held in June 2014, at which representatives from academia and other stakeholders (including other Research Councils) were invited to make inputs. This session will describe the outcomes from the review, and explain how its recommendations are being taken forward.

Plenary 2 – Champion (President’s Invited) Lecture

Tuesday 8 September 5pm – 6pm

Statistical Problems in Individualized Health

Scott Zeger

John Hopkins Bloomberg School of Public Health

William Osler, the founding physician of Johns Hopkins Medicine, said a century ago: “Variability is the law of life, and as no two faces are the same, so...no two individuals react alike and behave alike under the abnormal conditions which we know as disease” (Silverman, Murray & Bryan 2008). Osler could not have imagined the current explosion of health information enabled by the twin bio- and information technology revolutions that statistical scientists now use to explain that variation.

This talk will discuss Bayesian hierarchical models for longitudinal data as a framework for using population data and prior knowledge to address questions that patient's and population health managers ask. The methods will be illustrated with recent examples in which novel measurements and/or expert opinion inform health decisions.

5.1 Contributed – Medical – Survival Models

Wednesday 9 September 9am – 10am

Survival models for highly clustered censored data: Accurate inference based on integrated likelihoods

Giuliana Cortese, Nicola Sartori
University of Padua

In studies that involve censored time-to-event-data, stratification is frequently encountered due to several reasons, such as stratified sampling in multicenter clinical trials. Often the main interest is not in the clustering variables and the cluster-related parameters are treated as nuisance. When inference is about a parameter of interest in presence of many nuisance parameters, standard likelihood methods often perform very poorly and may lead to severe bias. This problem is particularly evident in models for clustered data with cluster-specific nuisance parameters, when the number of clusters is high with respect to the within-cluster size (Sartori, 2003, *Biometrika*). However, it is still unclear how the presence of censoring would affect this issue.

We consider clustered failure time data under independent censoring, and propose frequentist inference based on an integrated likelihood (Severini, 2010, *Biometrika*). This pseudo likelihood provides very accurate inference in presence of many nuisance parameters or small sample sizes. We show an application of the proposed method in stratified Weibull models. A data example about late-stage HIV-infected patients is used to compare the new approach with the existing alternatives, such as frailty models and stratified Cox's models.

Simulation studies show that appropriately defined integrated likelihoods provide very accurate inferential results in all circumstances, such as for highly clustered data or heavy censoring, even in extreme settings where standard likelihood procedures lead to strongly misleading results. We show that the proposed method performs generally as well as the frailty model, but it is superior when the frailty distribution is seriously misspecified.

5.1 Contributed – Medical – Survival Models

Wednesday 9 September 9am – 10am

Outlier Identification in Robust Joint Models

Lisa McCrink

Queen's University Belfast

To better understand the complex relationship between changing longitudinal responses over time and the impact on survival, joint modelling techniques are recommended to simultaneously analyse both types of responses. This is particularly useful in medical applications where it is common to collect individuals' changing blood measurements over time alongside time-to-event information. Due to the obvious potentials for such techniques, there has been a dramatic increase in the volume of literature surrounding joint models, however few account for the detrimental effect of longitudinal outliers.

This research extends recent work which has investigated and verified the need to use more robust approaches to accommodate longitudinal outliers through t-distributional assumptions. Alongside the improvement in accuracy and precision of parameter estimation through the use of more robust approaches, one of the key benefits to robust joint modelling is in the identification of outlying individuals, which is the focus of this research.

Through a simulation study, the predictive accuracy of robust joint models is clearly demonstrated under various conditions including the proportion and extremity of outliers present. The practical implications of such conclusions are investigated through an illustrative example which explores the impact of changing blood measurements on the survival of end stage renal disease patients. The potential use of such techniques is illustrated through the identification of outlying renal patients who are shown to have significantly worse survival than typical patients. The classification of renal patients as outliers has the potential therefore to assist clinicians in the improvement of survival rates for such patients.

5.1 Contributed – Medical – Survival Models

Wednesday 9 September 9am – 10am

Methods for analysing recurrent events in time-to-event (survival) data: an application in Hospital Episode Statistics

*Annie Herbert, Ruth Gilbert, Arturo González-Izquierdo, Leah Li
Institute of Child Health, University College London*

Time to recurrent events, such as hospital admissions, heart attacks, or machine failure, is rarely analysed using methods which consider events past the first. This is despite recurrent events being common in longitudinal research studies, and holding valuable information about the efficacy of operations or treatments. I explored several methods for analysing recurrent event data, including the Wei, Lin, & Weissfeld and Prentice, Williams, & Peterson models, to assess the effect on study conclusions when including events past the first event in analysis, which model gives a superior fit, and the best way to interpret such results. I will present my findings using an application of these methods to study recurrent admissions in Hospital Episode Statistics.

5.2 Contributed – Official – Census

Wednesday 9 September 9am – 10am

Plans and challenges for the 2021 Census in England and Wales

*Cal Ghee, Garnett Compton, Paul Waruszynski
Office for National Statistics, UK*

Following the Government's endorsement of the National Statistician's recommendation on 'the census and future provision of population statistics in England and Wales', the ONS Beyond 2011 Programme has been closed and replaced by the new Census Transformation Programme.

The new programme is focusing on developing the strategies and plans needed for delivery of the following major strands of work:-

1. An online census in 2021;
2. Integrated statistical outputs that make use of administrative data and surveys in conjunction with the census;
3. A recommendation for the future provision of population statistics beyond 2021.

This presentation will outline ONS plans for Strands 1 and 2: to deliver a predominantly digital census while making the most effective use of administrative and survey data in its design, operation and outputs.

It will cover the challenges of providing a census in 2021 that is 'digital by default', while building on the successes and lessons from the 2011 Census. Main areas that will be outlined include plans to address the challenge of digital exclusion while maximising the benefits of electronic data collection such as data quality, real-time response information and reducing processing time. Strand 2 is new for 2021, and looks at enhancing the traditional census building on the understanding of the opportunities and limitations of administrative data gained in Strand 3. Challenges include considering the most effective use of administrative and survey data in: optimising census data collection operations, estimating missing data, quality assuring results, reducing respondent burden or expanding topics covered.

5.2 Contributed – Official – Census

Wednesday 9 September 9am – 10am

Afghanistan: How to count the number of people in a conflict country

Kim Bradford Smith

UK Department for International Development

The paper will examine the political and technical issues in the Central Statistical Office of Afghanistan conducting a population census in a conflict country where no population census has been undertaken for over three decades, where population data is highly sensitive and contentious, and where people are suspicious of the government and government data. It will explain the 'compromise' method for conducting a large scale rolling survey over eight years, the logistical issues faced with undertaking the survey where ability is low and insecurity adds additional levels of complexity, and how insecurity and funding gap issues are dealt with. Finally it will discuss how the data is being used, and present options for aggregating key data/indicators across time and geography to estimate Afghanistan's total population and other key population indicators (such as mortality, fertility, life expectancy).

5.2 Contributed – Official – Census

Wednesday 9 September 9am – 10am

Electricity consumption and household characteristics: Implications for census-taking in a smart metered future

Ben Anderson, Sharon Lin, Andy Newing
University of Southampton

This paper assesses the feasibility of determining key household characteristics based on temporal load profiles of household electricity demand. It is known that household characteristics, behaviours and routines drive a number of features of household electricity loads and the roll out of domestic smart meters in the UK and elsewhere could enable the collection of high temporal resolution electricity monitoring data at the household level. Such data affords tremendous potential to invert the established relationship between household characteristics and temporal load profiles. Rather than use household characteristics as a predictor of loads, observed electricity load profiles, or indicators based on them, could instead be used to impute household characteristics. These micro level imputed characteristics could then be aggregated at the small area level to produce 'census-like' small area indicators. This paper briefly reviews the nature of current and future census taking in the UK before outlining the household characteristics that are to be found in the UK census and which are also known to influence electricity load profiles. It then presents descriptive analysis of two smart-meter like datasets of half-hourly domestic electricity consumption before reporting results from a multilevel modelling-based analysis of the same data. The paper concludes that a number of household characteristics of the kind to be found in UK census-derived small area statistics may be predicted from particular load profile indicators. The paper ends with a discussion of the steps required to test and validate this approach and the wider implications for census taking.

5.3 Contributed – Environment – Climate Change

Wednesday 9 September 9am – 10am

Sources of uncertainty in future projections of the carbon cycle

Alan Hewitt, Ben Booth, Chris Jones, Eddy Robertson, Andy Wiltshire, David Stephenson, Stan Yip
UK Met Office

The increased inclusion of carbon cycle processes within CMIP5 climate models provides a new framework to explore the relative importance of uncertainties in scenario and model representation to future land and ocean carbon fluxes. A two-way ANOVA approach, appropriate for an unbalanced design, was used to distinguish the relative importance of these uncertainties at different lead times.

For global ocean carbon fluxes, variance attributed to differences between RCP scenarios overtakes in importance the variance attributed to differences between models around 2030, completely dominating by 2100. This contrasts with global land carbon fluxes, where variance attributed to differences between models remain the largest uncertainty source beyond 2100. This suggests that modelled processes that determine ocean fluxes are currently better constrained than those of land fluxes, thus we can be more confident in linking different future socio-economic pathways to consequences of ocean carbon uptake than for land carbon uptake.

The apparent model agreement in atmosphere-ocean carbon fluxes, globally, masks strong model differences at a regional level. The North Atlantic and Southern Ocean are key regions, where differences in modelled processes represent an important source of uncertainty in projected regional fluxes. For the tropical land region, the divergence of scenario variance at the beginning of the century results from "anthropogenic land-use change". This emphasises how increased complexity in the representation of Earth system processes leads to a more nuanced picture of anthropogenic influence in future scenarios.

5.3 Contributed – Environment – Climate Change

Wednesday 9 September 9am – 10am

On constraining projections of future climate change using observations and simulations from multiple climate models

*Philip Sansom, David Stephenson, Thomas Bracegirdle
University of Exeter*

Appropriate statistical frameworks are required to make credible inferences about long term climate change based on projections from multiple climate models. Climate models are imperfect representations of the Earth system. The spread of projections of future climate simulated by different models represents a substantial source of uncertainty. The relationship between the models and the Earth system is also uncertain. A new hierarchical Bayesian framework is presented that is able to constrain projections of future climate using recent observations, while allowing for our uncertainty about the relationship between the models and the system.

Projections of future climate are also subject to many other sources of uncertainty. It is shown that if initial condition uncertainty in the simulations is not accounted for, then the estimated relationship between the historical and future states of the system may be biased. Observation uncertainty and natural variability in the actual system also play an important role when observations of the historical state are used to constrain the future.

The framework is applied to projections of surface temperature in the Arctic at the end of the century. The projected warming in some regions is reduced by more than 2K when constrained by historical observations. Observation uncertainty and natural variability can affect the estimated warming by up to 0.25K. The uncertainty about the climate response is reduced by up to 30% where strong constraints exist.

5.3 Contributed – Environment – Climate Change

Wednesday 9 September 9am – 10am

A Bayesian framework for verification and recalibration of ensemble forecasts: How uncertain is NAO predictability?

Stefan Siegert, David Stephenson, Philip Sansom, Adam Scaife, Rosie Eade, Alberto Arribas
University of Exeter

Predictability estimates of ensemble forecasting systems are prone to sampling uncertainty due to limited numbers of past forecasts and limited ensemble sizes. To account for such uncertainty, this study exploits a statistical signal-plus-noise framework to model observations and corresponding ensemble forecasts. The framework is formulated as a latent variable model, which is argued to be a believable parsimonious representation of many forecast ensembles being developed for climate services. The posterior distribution of the framework parameters is inferred by Markov Chain Monte Carlo methods. The framework can be used to mimic the possible behaviour of the forecasting system and observations over new time periods and/or with different ensemble sizes. In addition to quantifying skill uncertainty, it is demonstrated how the framework can be used to construct well-calibrated predictive probability distributions. The framework is used to address important questions concerning the skill of winter hindcasts of the North Atlantic Oscillation for 1992-2011 issued by the Met Office GloSea5 climate prediction system. Although there is much uncertainty in the correlation between ensemble mean and observations, there is strong evidence of skill: the 95% credible interval of the correlation coefficient of [0.19,0.68] does not overlap zero. There is also strong evidence that the forecasts are not exchangeable with the observations: With over 99% certainty, the signal-to-noise ratio of the forecasts is smaller than the signal-to-noise ratio of the observations, which strengthens previous findings that raw forecasts should not be taken as representative scenarios of the observations, and that post-processing of the forecasts is necessary.

5.4 Contributed – Medical – Biostatistics

Wednesday 9 September 9am – 10am

Translational Statistics, Dynamic Nomograms and Mean Residual Life

*John Newell, Alberto Alvarez-Iglesias, Amirhossein Jalali, John Hinde
NUI Galway*

Translational Medicine promotes the convergence of basic and clinical research disciplines and the transfer of knowledge on the benefits and risks of therapies. It aims to improve the flow from laboratory research through clinical testing and evaluation to standard therapeutic practice. This transfer of knowledge informs both clinicians and patients of the benefits and risks of therapies.

In an analogous fashion we propose the concept of Translational Statistics to facilitate the integration of Biostatistics within clinical research and enhance communication of research findings in an accurate manner to diverse audiences (e.g. policy makers, patients and the media).

In survival analysis for example, the mean residual life function provides a clear and simple summary of the effect of a treatment or a risk factor in units of time, avoiding hazard ratios or probability scales, which require careful interpretation. Estimation of the mean residual life is complicated by the upper tail of the survival distribution not being observed due to right censoring. In a recent publication, Alvarez-Iglesias et al (2015) presented a novel semi-parametric method that combines existing non-parametric methods and an extreme value tail model, where the limited sample information in the tail (prior to study termination) is used to estimate the upper tail behaviour.

We will consider also how modern web-based computing allows the simple development of interactive dynamic tools for communicating and exploring research findings. Various examples will be used to illustrate the implementation and application of these ideas.

5.4 Contributed – Medical – Biostatistics

Wednesday 9 September 9am – 10am

Spatial and temporal signatures of dengue in Singapore

Sourav Das, Alex Cook, Shi Yuan
University of Bristol

Dengue infects over ten to twenty thousand individuals in the city state of Singapore each year. Prevention, in the absence of a vaccine, amounts to vector control, costing around \$50m per annum. To help deploy control efforts more efficiently, we have analysed the spatio-temporal patterns of over 80 000 notified dengue cases in Singapore over a period of ten years for whom date of onset and home address are known. Using spatial point process methodologies and Monte Carlo, we identify characteristic features of epidemic and endemic dengue within the equatorial city state and its second order spatial signature. We find that despite Singapore's uniformity of urban landscape the incidence of dengue is heterogeneous and the strength of the association with mosquito breeding intensities varies from epidemic to non-epidemic periods. We also observe that the spatial and temporal variation influencing the transmission and prevalence of the epidemic are separable.

5.5 Contributed – Methods & Theory – Historical Review

Wednesday 9 September 9am – 10am

The Information Imperative in Fisher's Statistics

John Aldrich

University of Southampton

This paper examines the history of some of the most important concepts in statistical inference, viz. information, sufficiency, efficiency, ancillarity and maximum likelihood. These concepts were originally linked to the information imperative in Ronald Fisher's conception of statistics: the statistician's task is "the reduction of data" so as "to isolate the whole of the relevant information contained in the data." In 1921 Fisher thought the imperative meaningful because a sufficient statistic always existed and because it could be found by the method of maximum likelihood. Fisher soon discovered that this original conception was flawed and subsequently re-specified the objective as minimising the loss of information and as fully recovering the information lost. He never claimed that the new imperatives could be realised but he judged them promising: in 1925 he wrote that his latest work "seems to make possible" a theory with "some approach to completeness" while developments in 1934 led him to judge that the treatment of the general problem is "ripe for solution." Later Fisher tried to come to terms with the realisation that the imperatives could not be realised. The information project was almost entirely Fisher's work. Critics judged the amount of information "too complicated and remote to serve as a principle" unlike the "frequency of errors in judgement" which was a "sufficiently simple and unquestionable principle." Neyman's judgement (of 1935) has been followed in the modern theory of inference where Fisher concepts have a role but not the primary one for which they were originally introduced.

5.5 Contributed – Methods & Theory – Historical Review

Wednesday 9 September 9am – 10am

Multiple regression, longitudinal data and welfare in the nineteenth century: reflections on Yule (1899)

Ian Plewis

University of Manchester

G. Udny Yule served as President of the Society between 1924 and 1926. Twenty five years earlier he read a paper on the causes of changes in pauperism which is a landmark in social statistics. It was the first time that the multiple regression method had been applied to social science data and Yule applied it in a way that will be familiar to today's quantitative social scientists. Another remarkable feature of Yule's 1899 paper is that it is an analysis of longitudinal data. He also linked Census data to administrative data in a way that seems quite modern. These data have been reconstructed from Poor Law returns etc. and then reanalysed; the findings from these reanalyses are presented here for the first time. The presentation briefly considers the social context of Yule's paper and its relevance to current methodological and substantive debates about the analysis of observational data and about the operation of the welfare system. I then outline the strengths and weaknesses of Yule's analysis before going on to show how twenty first century social statisticians might use random and fixed effects approaches to answer Yule's main research question in the light of the data available to him in the 1890s. These reanalyses indicate that Yule's main statistical finding about pauperism stands up but that the effects are much more heterogeneous than he was able to elucidate, and the interpretation of those effects much less straightforward.

5.5 Contributed – Methods & Theory – Historical Review

Wednesday 9 September 9am – 10am

Why non-parametric estimation of the ROC curve is hard.

Matt Stapleton

AWE

The Receiver Operating Characteristic (ROC) curve is a popular tool for characterising the performance of classifiers. It illustrates the trade-off between false positives and false negatives inherent in the design of any classifier by providing the sensitivity for any given specificity. However, the true ROC curve is rarely known in practice and is typically estimated from data, leading to the empirical ROC curve. There is considerable literature on methods for generating such estimates, covering the full spectrum of parametric, semi-, and non-parametric techniques. Non-parametric estimation is particularly straightforward, yet generating reliable confidence intervals from small samples remains a formidable theoretical challenge.

We make use of the theory of order statistics to calculate sampling properties of the empirical ROC curve. These highlight the theoretical challenges involved in non-parametric estimation of the ROC curve, in particular the sensitivity at fixed specificity for small samples. Our approach also provides an alternative interpretation of the ROC curve itself, which calls into question the usual approach of generating confidence intervals at fixed thresholds based on the binomial distribution. Our aim is to provide insight that will be useful to both researchers and practitioners in this area.

5.6 Invited – Prize Winners RSC 2015

Wednesday 9 September 9am – 10am

Analysis of Linguistic Change in Glaswegian Dialect

Craig Alexander

University of Glasgow

The Sounds of the City project looks into speech sounds of Glasgow past and present. Just as the landscape of the city has experienced transformations over time, the speech of its residents has been continuously changing. We discuss the use of chain graph models as a method of static analysis on specific aspects of language using data from the Sounds of the City project. The chain graph model is proposed to provide a simple visual description of the dependency structure within the data. The model is constructed by fitting a linear mixed model with a multivariate response to account for the variation between individual speakers. Future research will consider tracking temporal dependency in speech with the use of a hidden Markov model (HMM), with the unobserved variable being introduced to track the temporal dependency.

5.6 Invited – Prize Winners RSC 2015

Wednesday 9 September 9am – 10am

Difficulties with popular calibration techniques.

*Gareth Davies, Jonathan Gillard, Anatoly Zhigljavsky
Cardiff University*

The problem of calibration in survey sampling can be considered as a convex optimization problem. This talk will outline various choices of the cost function used as part of this optimization problem. We begin by considering classical functions including the quadratic, raking and logit functions. The benefits and drawbacks of these classical calibration functions shall be explored. The main criticism of the quadratic and raking functions is that they can result in negative and/or extreme survey weights. In this talk, we shall introduce a new class of calibration functions that guarantee the calibration weights are bounded in a pre-specified interval.

5.7 Contributed – Data Science – Miscellaneous

Wednesday 9 September 9am – 10am

A mixed model approach to the relationship between molecular structure and biological activity: application to high-throughput screening of potential new medicines

Nicholas Galwey, Steven Barrett, Gonzalo Colmenarejo
GlaxoSmithKline R&D

Large sets (>500K) of chemical compounds can be tested in a biological assay to identify chemical starting points for the development of new medicines, in a process called high-throughput screening (HTS). The utility of such a HTS measure is assessed by the accuracy with which it predicts the values of a more intensive 'gold standard' measure on the same compounds. It is possible that the accuracy can be increased if each compound is assessed not only on its own HTS value, but also on those of compounds with similar molecular structures, by appropriately modelling the local structure-activity relationship around the compound in question. This presentation introduces a mixed-model approach to such 'borrowing' of information. Molecular-structure similarities provide the covariance matrix for the model-fitting process. To reduce the computational burden, standard mixed-model software tools are not used: instead, singular value decomposition is performed on the covariance matrix, the eigenvectors are used as predictors in a linear regression model, and the eigenvalues are used to determine the number of clusters to which the pattern of similarities among the compounds is equivalent. In the special case where the HTS values are in fact replicate observations on a smaller number of compounds, this method gives the shrunk mean for each compound. On the data sets studied so far, this approach gives a small but consistent increase in the accuracy of prediction of the 'gold standard' measure, raising the correlation coefficient between the measures from $r = 0.637$ to $r = 0.700$.

5.7 Contributed – Data Science – Miscellaneous

Wednesday 9 September 9am – 10am

The chopthin algorithm for resampling

Axel Gandy, F. Din-Houn Lau
Imperial College London

Resampling is a standard step in particle filters and more generally sequential Monte Carlo methods. We present an algorithm, called chopthin, for resampling weighted particles. In contrast to standard resampling methods the algorithm does not produce a set of equally weighted particles; instead it merely enforces an upper bound on the ratio between the weights. A simulation study shows that the chopthin algorithm consistently outperforms standard resampling methods. The algorithm chops up particles with large weight and thins out particles with low weight, hence its name. It implicitly guarantees a lower bound on the effective sample size. The algorithm can be implemented very efficiently, making it practically useful. We show that the expected computational effort is linear in the number of particles. Implementations for C++, R (on CRAN) and for Matlab are available

5.7 Contributed – Data Science – Miscellaneous

Wednesday 9 September 9am – 10am

Probabilistic uncertainty analysis of energy system models, with an application to a model for the UK energy market

*Amy Wilson, Michael Goldstein, Chris Dent
Durham University*

Computer models are used in the study of energy systems to simulate real-world processes. These models take a set of inputs, which are often uncertain, and from them compute some output of interest. Results from model evaluations could be used to drive changes in energy policy or to ensure the future security of energy supply so it is important to have a thorough understanding of any uncertainty associated with model outputs; such uncertainty may stem from uncertain inputs or uncertain system properties. For complex models, using Monte Carlo simulation to investigate uncertainty can be infeasible due to the time taken to perform model evaluations.

The Dynamic Dispatch model (DDM) is an energy systems model commissioned by the Department for Energy and Climate Change. The DDM was used to estimate the amount of power generation that the government should procure in a capacity market auction in 2014 to meet electricity demand requirements in 2018. Many of the inputs to the DDM were uncertain, such as future changes in electricity demand.

The DDM will be used to demonstrate how Gaussian Process emulators can help quantify uncertainty in energy system models. An emulator was constructed for the DDM based on a limited number of publicly available model evaluations, which did not effectively cover the feasible input space. Some of the statistical issues arising from this will be discussed. Results will also be presented from an uncertainty analysis of the DDM, investigating the effect of input uncertainty on the output.

6.1 Invited – Medical – Papers from the RSS Journal

Wednesday 9 September 10.05am – 11.25am

Modelling reporting delays for outbreak detection in infectious disease data

Angela Noufaily, Yonas Ghebremichael-Weldeselassie, Doyo Gragn Enki, Paul Garthwaite, Nick Andrews, Andre Charlett, Paddy Farrington
University of Warwick

The delay that necessarily occurs between the emergence of symptoms and the identification of the cause of those symptoms affects the timeliness of detection of emerging outbreaks of infectious diseases, and hence the ability to take preventive action. We study the delays that are associated with the collection of laboratory surveillance data in England, Wales and Northern Ireland, using 12 infections of contrasting characteristics. We use a continuous time spline-based model for the hazard of the delay distribution, along with an associated proportional hazards model. The delay distributions are found to have extremely long tails, the hazard at longer delays being roughly constant, suggestive of a memoryless process, though some laboratories appear to stop reporting after a certain delay. The hazards are found typically to vary strongly with calendar time, and to a lesser extent with season and recent organism frequency. In consequence, the delay distributions cannot be assumed to be stationary. These findings will inform the development of outbreak detection algorithms that take account of reporting delays.

6.1 Invited – Medical – Papers from the RSS Journal

Wednesday 9 September 10.05am – 11.25am

Bayesian analysis of measurement error models using integrated nested Laplace approximations

*Stefanie Muff, Andrea Riebler, Leonhard Held, Havard Rue, Philippe Saner
University of Zurich*

Important explanatory variables in applied sciences are often difficult to measure and thus may contain considerable measurement error (ME). If such ME is ignored, parameter estimates and confidence intervals in statistical models often suffer from serious biases. Bayesian inference of ME problems is usually based on hierarchical models and has so far required the use of MCMC techniques. Here, we show how Gaussian ME models fit into the framework of latent Gaussian models in the context of generalized linear mixed model regression. We present an extension of the Integrated nested Laplace Approximations (INLA) approach to the classical and the Berkson ME models, which are suitable for continuous error-prone covariates. Several examples illustrate the applicability of the new models, and further extensions are discussed.

6.5 Invited – New directions in statistical methods for environmental exposure and health

Wednesday 9 September 10.05am – 11.25am

Modelling nutrient dynamics in aquatic ecosystem: their response to storm-driven events

*Monica Pirani, Anouska Panton, Duncan A. Purdie, Sujit K. Sahu
S3Ri, University of Southampton*

Nutrients, primarily nitrogen and phosphorus based compounds, are aquatic pollutants that are harmful to human health and ecosystems. These pollutants are driven by river currents and show dynamics that are heavily affected by storms. As a result the nutrient budget in the receiving estuaries and lakes change suddenly and seasonally causing ecological damage to resident wild life and fish population.

This paper proposes a statistical change-point model that interacts between seasons and storm events to capture the nutrient dynamics enabling us to produce accurate estimates of the fluxes and their annual deposited levels. In particular, we present a study on nutrient dynamics on the Hampshire Avon River, which flows to the southern coast of the UK. We model daily time series data for more than a year during 2013-14 in which period there were multiple severe storm events, which were not seen in the recent past. Adopting a non-linear hierarchical model within the Bayesian inference framework we quantify riverine macronutrient fluxes and the model also allows us to estimate past fluxes based on input river flow values. Out of sample empirical validation methods justify our approach, which captures also the dependencies of pollutant levels with waterbody characteristics such as temperature and conductivity. The model is easily implemented in publicly available software, which makes it an ideal candidate for similar data modelling and analysis problems for estuarine pollution.

6.5 Invited – New directions in statistical methods for environmental exposure and health

Wednesday 9 September 10.05am – 11.25am

The Regression Discontinuity Design in Policy Evaluation: Protective Effects of the Smoke Free Legislation on birth outcomes.

*Ioannis Bakolis, Nicky Best, Daniela Fecht, Anna Hansell, Christopher Millett, Kevin Garwood, Paul Elliot, Susan Hodgson
Imperial College London*

Causal inference is a central goal of epidemiology. Recent methodological developments have provided support for the focus on isolating factors that can be seen as causes of a specific health outcome, and there is an increasing understanding that observational studies should, where possible, be designed to approximate randomized experiments. In this context, we employed a novel quasi-experimental design, entitled regression discontinuity (RD) design, which has not yet been widely applied to epidemiological research. We aim to provide the theoretical framework of RD design for the analysis of routinely collected data (Hospital Episodes Statistics (HES)) to quantify the effect of smoke-free legislation on adverse birth outcomes in England. The RD design was used to assess the discontinuity in the expected risk of the birth outcomes, which were visually inspected by graphical analysis. Previous studies have employed interrupted time series designs to explore changes in risk of adverse pregnancy outcomes before and after implementation of smoke-free legislation. However, standard interrupted time series designs could potentially lead to biased effect estimates due to unobservable confounders. The RD design guarantees comparability between pregnancies lying on either side of the cut-off point so this bias could be potentially eliminated. Smoke-free legislation in England is likely to have affected pregnant women's exposure to tobacco smoke and environmental tobacco smoke with consequent impacts on pregnancy outcomes. RD design is a novel approach which could effectively quantify this exposure-health effect.

6.5 Invited – New directions in statistical methods for environmental exposure and health

Wednesday 9 September 10.05am – 11.25am

Coherent public health forecasts for environmental risk assessment

Vasilis Kontis, James E Bennett, Colin D Mathers, Majid Ezzati
Imperial College London

Estimates of future mortality are a needed component to analyses of the impacts of environmental and health risk factors, which are often estimated as a proportion of disease-specific death rates. Our objective is to develop a methodological and computational approach to obtain coherent mortality forecasts at the national and multi-country levels.

We use a Bayesian model averaging approach to forecast all-cause mortality and life expectancy. Our ensemble consists of models that incorporate important features of mortality time series, such as similarities in death rate levels and trends in adjacent age groups and birth cohorts, and the fact that recent data may be more informative for forecasting than earlier data. The performance of individual models is assessed using external (out-of-sample) validity tests. Based on their forecasting accuracy, models are assigned weights and combined into a single model average. The large number of runs required is manageable thanks to the INLA (Integrated Nested Laplace Approximation) platform.

Preliminary results show that the best performing model differs across countries and sexes, and that the predictive performance of individual models varies significantly depending on the length of the forecasting horizon and the time period used for validation. Bayesian model averaging provides a method to obtain forecasts that are more consistent and reliable, compared with using a single model.

6.5 Invited – New directions in statistical methods for environmental exposure and health

Wednesday 9 September 10.05am – 11.25am

Modelling geo-located public health data using spatio-temporal log-Gaussian Cox processes

Theresa Smith, Peter Diggle, Ben Taylor
Lancaster University

We present a spatio-temporal log-Gaussian Cox process (LGCP) and a Bayesian approach to joint estimation of the latent Gaussian process (GP) and the effects of environmental predictors on the spatio-temporal intensity surface. LGCPs are a type of non-homogenous Poisson point process where the log intensity surface is a GP. A point process approach is appropriate when each observation is indexed to a particular point in space and time. This in contrast to the common area-level approach in epidemiology where observations and risk factors are summarized over several small regions (e.g., counties or local authorities). The spatially-continuous approach inherent in LGCPs more naturally accommodates risk factors measured on different spatio-temporal units and avoids some forms of ecological bias. We compare maximum likelihood and Bayesian techniques for estimating systematic trends in the spatio-temporal risk surface as well as the latent Gaussian process. Finally we use a spatio-temporal LGCP to investigate the roles of environmental and socio-economic risk-factors on the incidence of *Campylobacter* (a common bacterial cause of food borne disease) in the United Kingdom.

6.6 Invited – Data Science – Urban Informatics

Wednesday 9 September 10.05am – 11.25am

Detecting trends and changes in urbanization via statistical modelling of land use maps

Massimo Ventrucchi, Daniela Cocchi, Marian Scott
University of Bologna

Land use maps are a powerful resource to study the dynamics of urban growth. Official land use data, routinely produced by a wide range of institutions, are usually the result of processing remote sensing images. After complex elaborations, maps are provided to the public in two possible spatial formats: vector and raster. Vector maps are a collection of polygons, where each polygon belongs to a land use category. Raster maps result from converting polygons into a regular lattice or grid, where each grid box is assigned a land use category. Visual inspection of raster maps can immediately provide important information on the urban distribution over space, however, formal methods for the detection of spatio-temporal trends in this type of data are required. Modelling the spatial dependencies using the traditional Matern class of covariance functions is problematic because of the high dimensionality of these maps. We discuss efficient P-spline smoothing models for raster maps, which give computational advantages and therefore allows urbanization trends to be investigated, even in relatively large regions such as a city or a metropolitan area. An application of the proposed models is illustrated on urban data on the metropolitan area around Bologna, Italy. One of the most challenging application goals is the detection of regions showing a significant change in the urbanisation process over time.

6.6 Invited – Data Science – Urban Informatics

Wednesday 9 September 10.05am – 11.25am

Communicating Big Urban Data - Smart Cities, Realtime Feeds, Simulation and The Crowd

Andrew Hudson-Smith
University College London

The paper explores the rise of big data in the context of Smart Cities. It explore how contributed data from systems such as the London Oyster Card or scrapped data from bike share schemes are opening up new insights into the city. The paper explores new methods for collecting data, specifically social network data and how these can be visualised in terms of viewing the life of the city. Finally it explores the move towards realtime data and the link to urban simulation to provide an insight into the next 30 minutes of a urban system. It concludes with a look at the rise of the Internet of Things and how emerging test beds are allowing data to be communicated in new and evolving ways.

7.1 Invited – Medical – Synthesizing evidence

Wednesday 9 September 11.50am – 1.10pm

Individual participant data meta-analysis: opportunities and challenges

Richard Riley
Keele University

Objectives

Individual participant data (IPD) meta-analysis involves obtaining and then synthesising the raw, individual-level data from primary studies [1]. It is often considered the gold-standard approach to evidence synthesis, as it has many potential advantages over traditional methods that only use aggregate data (such as effect estimates and confidence intervals). In this talk I demonstrate a range of important applications of IPD meta-analysis that are either inappropriate or not feasible when only aggregate data are used

Methods

Applications will include: (i) stratified medicine, where a treatment-covariate interaction is of interest; (ii) risk prediction, where a prognostic model needs both development and external validation; and (iii) epidemiological research, where a potential risk factor is examined.

Results

The examples illustrate that (i) treatment-covariate interactions estimated using IPD may differ dramatically to those estimated using aggregate data; (ii) IPD allows a prediction model to be examined across multiple studies, to identify heterogeneity in its calibration and discrimination performance, and to identify whether recalibration is needed in particular populations; and (iii) IPD does not solve all the issues for meta-analysis in epidemiology, but facilitates novel multivariate meta-analyses that limit missing adjustment factors in some studies.

Conclusions

IPD meta-analysis is an increasingly feasible approach to evidence synthesis, and should be increasingly used. It especially encourages a move toward stratified medicine by improving the identification of predictors of treatment response, risk factors, and reliable prognostic models.

[1] Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of IPD: rationale, conduct, and reporting. *BMJ* 2010;340:c221

7.1 Invited – Medical – Synthesizing evidence

Wednesday 9 September 11.50am – 1.10pm

Network meta-analysis of multiple outcomes

Sofia Dias

University of Bristol

Objectives: Network meta-analysis (NMA) is now well established as a method for deciding which treatment is most effective or cost-effective, out of several alternatives. NMA is used routinely in NICE Technology Appraisals and Clinical Guidelines.

Often, in addition to multiple treatments, several measures are used to assess patients' response to treatment (outcomes), and these can be reported in different formats. Separate analyses of each outcome, typically using different randomised controlled trials, can lead to conflicting conclusions. There is therefore a need to extend NMA to multiple outcomes, so that decisions can be based on a coherent model of treatment differences on all relevant outcomes.

Methods: The Bayesian NMA modelling framework and assumptions will be described. Models incorporating within-trial correlations (ie in the likelihood) and outcome relationships, based on what is known from the data generating process, will be presented for some key examples.

Results: Outcome relationships can often be incorporated in the NMA and have the advantage of producing joint relative effect estimates which appropriately account for the correlation in the outcomes.

Conclusions: Different applications have different types of evidence structure, which may require a different model. However, using a joint model for all the outcomes, and adequately incorporating what is known about the relations between different relative effects is essential for making decisions based on multiple outcomes and should be incorporated in the standard NMA framework.

A framework for evaluating and communicating the, sometimes strong, assumptions made to clinicians and decision makers is required.

7.1 Invited – Medical – Synthesizing evidence

Wednesday 9 September 11.50am – 1.10pm

Using Real World Evidence (RWE) in Health Technology Assessment (HTA) Decision Making - Challenges & Potential

Keith Abrams

University of Leicester

In HTA decision making much emphasis is often put upon evidence from Randomised Controlled Trials (RCTs). However, both changes in the way RCTs are undertaken (especially for pharmaceutical regulation) and a desire for earlier HTA decisions, means that such evidence will become increasingly more limited in both nature and scale - for example smaller (possibly adaptive) trials with shorter follow-up and using surrogate endpoints. Thus, other sources of evidence will be required in order that decisions can be made in a timely manner. The use of RWE has been advocated as a way of addressing some of these issues, but it is not without its challenges. These will be illustrated, using recent examples from National Institute for Health & care excellence (NICE), by considering how RWE can be used in estimating Relative Effectiveness (RE) by augmenting RCT evidence, including within a Network Meta-Analysis (NMA) setting and mapping between outcomes. Finally, other potential areas for the use of RWE in HTA will be discussed as well as the ongoing Innovative Medicines Initiative (IMI) GetReal project which is evaluating the potential use of RWE in HTA.

7.2 Invited –Official – The Science of Electoral Forecasting

Wednesday 9 September 11.50am – 1.10pm

Stephen Fisher
University of Oxford

An exit poll for the 2015 UK general election predicted that the Conservatives would emerge as the clear winners despite pre-election opinion polls suggesting a much closer race. This was the third outing for an exit poll methodology based on the principle of revisiting previous exit poll sampling locations and analysing variation in the change between exit polls at the constituency level. Statistically this was the least accurate of the three and it importantly failed to predict that the Conservatives would emerge with an overall majority. But politically it was the most profound in revising expectations of the likely outcome towards a much more accurate impression. This paper reviews the methodology and experience of conducting and the exit poll and discusses potential lessons for future exit polls.

Andreas Murr
University of Oxford

Who do you think will win in your constituency? Most citizens correctly answer this question, and groups of citizens are even better at answering it. Combining individual forecasts results in the "wisdom of crowds" explained by Condorcet's jury theorem. This paper demonstrates the accuracy of citizen forecasts in seven previous British General Elections between 1964 and 2010, and reports what citizens interviewed in February and March forecasted for the election in May 2015. The combination of citizen forecasts predicts who wins in a constituency and by how much, plus how many seats a party will win and who becomes Prime Minister. Citizen forecasts are direct, accurate, and comprehensible, and so pollsters should collect them and communicate their results more often.

Ben Lauderdale
LSE

We describe a method for forecasting British general elections that integrates both national and constituency polling to provide seat forecasts and provide consistent probabilistic forecasts across seat and national-level outcomes. The model was overconfident in its national-level seat predictions in 2015 because of the large national polling miss and Conservative over-performance in marginal seats, only the first of which was captured in the model. We discuss lessons for forecasting future election-forecasting projects, particularly in situations where calibration data is very limited and correlated errors are likely.

7.5 Invited – Modern methodological challenges in statistics

Wednesday 9 September 11.50am – 1.10pm

Nonparametric Network Summaries

Sofia Olhede, Patrick Wolfe, Pierre Andre Maugis
University College London

Relational data have become an important component of modern statistics. Networks, and weighted networks, are ubiquitous in modern applications such as disease dynamics, ecology, financial contagion, and neuroscience. The inference of networks is harder, in parts because the measure placed on the observables need to satisfy sets of permutation invariances, and most networks are very sparse, with most possible relations not present.

This talk will explore how to best construct nonparametric summaries of such objects, in such a way that the underlying statistical model of the observations is well described, and any estimators computable with scalable algorithms.

7.5 Invited – Modern methodological challenges in statistics

Wednesday 9 September 11.50am – 1.10pm

Random projection ensemble classification

Tim Cannings, Richard Samworth
University of Cambridge

We introduce a very general method for high-dimensional classification, based on careful combination of the results of applying an arbitrary base classifier to random projections of the feature vectors into a lower-dimensional space. In one special case that we study in detail, the random projections are divided into non-overlapping blocks, and within each block we select the projection yielding the smallest estimate of the test error. Our random projection ensemble classifier then aggregates the results of applying the base classifier on the selected projections, with a data-driven voting threshold to determine the final assignment. Our theoretical results elucidate the effect on performance of increasing the number of projections. Moreover, under a boundary condition implied by the sufficient dimension reduction assumption, we show that the test excess risk of the random projection ensemble classifier can be controlled by terms that do not depend on the original data dimension. The classifier is also compared empirically with several other popular high-dimensional classifiers via an extensive simulation study, which reveals its excellent finite-sample performance.

7.5 Invited – Modern methodological challenges in statistics

Wednesday 9 September 11.50am – 1.10pm

Contemporary challenges in changepoint analysis

Idris Eckley
Lancaster University

In recent years there has been a surge of interest in the development of new methods for tackling changepoint problems. Much of this work has focused on the analysis of univariate time series, with contributions in this area having an impact in various disciplines, notably climate science and genomics. However with the increased use of low-cost, multivariate sensors, there is a growing need to be able to efficiently and accurately identify changes in structure within multivariate time series. In moving to this more complex setting, the problem of detecting changepoints becomes more subtle. We will outline some recent developments in this area motivated by some ongoing work with industrial collaborators.

7.6 Invited – The rise of data science

Wednesday 9 September 11.50am – 1.10pm

The statistical face of data science: what will future graduates need?

David Hand

Imperial College London

Data science is an ill-defined term, spanning both data manipulation and inference. The disciplines which deal with these domains are, respectively, computer science and statistics. This means that data science is fundamentally interdisciplinary, and in particular that teaching in universities must be a collaborative effort. To explore the impact of the rise of data science on statistics academia, I look at who is being taught, how they are being taught, what they are being taught, and what they should be taught.

7.6 Invited – The rise of data science

Wednesday 9 September 11.50am – 1.10pm

Data science - a commercial perspective

Gordon Blunt

Gordon Blunt Analytics Ltd

The term 'data science' has become common in recent years. However, it seems that this is little more than the latest jargon that has become attached to a discipline which is fundamentally statistical. What distinguishes 'data science' from 'traditional statistics' is the size of modern data sets, which could have hundreds of millions, or billions, of observations.

Some of the work needed in data analysis has become more complex because of such 'big data'. For example, checking for errors is a lot harder when there are 100 million cases rather than just 100. The latter can be done with paper and pencil, but the former cannot be done without the aid of a computer. The skills that have been taught for decades, such as exploratory data analysis, inference and so on are still needed, and, I would argue, are even more important in the era of massive data sets.

One complication is that, in the commercial sector, clients want answers; they are usually less concerned with methods than in having timely and useful answers to their problems. Communication and the ability to work with non-statisticians are likely to be just as important skills as statistical ability.

Given this, I will describe the skills that, in my experience, are needed by a statistician working in the commercial sector, and hence suggest some ideas about what could be taught in statistics courses.

7.6 Invited – The rise of data science

Wednesday 9 September 11.50am – 1.10pm

Data Science and Education

Sofia Olhede, Patrick Wolfe
University College London

Data Science is emerging as one of the key growth areas in society. With an ever-growing range of possible Data Science applications, with innovations of usage of Data Science in education itself (so-called learning analytics), the need for trained data analysts seems boundless.

Despite this clear need, it is less clear how these analysts are to be trained. Clearly a strong foundation in computation is needed, a considerable understanding of sampling, and overall population variability as well. A Data Scientist is not quite a statistician, but neither is she a computer scientist. Data science also interacts with various application fields, where to fulfil its promise, and live up to the hype, a number of challenges have to be faced. This seems to put an almost impossible requirement on the desirable skills of the data scientist.

The overarching question linked to that of how to train a Data Scientist, is whether Data Science will eventually form its own field, with its own core training and its own professional society. I will discuss what needs to happen for this to become a reality, and whether the area is about to stand on its own two feet.

Plenary 3 – Significance Lecture

Wednesday 9 September 2.10pm – 3.10pm

Politics is broken: How can we measure, analyse and understand increasingly fragmented electorates?

Alberto Nardelli

Data Editor, Guardian News & Media

Using the failure of the pre-election polls as a jumping off point, the lecture will explore - through data - the ways in which the UK electorate is changing, how those changes are reflected across Europe, and why those changes make it difficult to accurately predict how people will vote.

Key trends to be discussed include:

- The rise of nationalism - both positive and negative
- The crisis of the moderate left, and how it is struggling to find its place in a post-recession world of austerity politics
- The fragmentation of party support: how new and smaller parties are gaining in popularity and what this means for the 'establishment' parties

The lecture will conclude by highlighting the challenges for statisticians and data journalists:

- Statistical models use past behaviour as a reference point for predicting future behaviour. But is the past a useful guide when the present is so different from what came before?
- Reporting the numbers is not enough. How can we get better at putting data in context, at seeing the underlying trends, and understanding what the numbers mean at a human level?

8.1 Contributed – Medical – Modelling case studies

Wednesday 9 September 3.15pm – 4.15pm

Acute kidney injury amongst chronic kidney disease patients: a case-study in statistical modelling

Ozgur Asar, James Ritchie, Philip Kalra, Peter Diggle
Lancaster University

Chronic kidney disease (CKD) and acute kidney injury (AKI) are two important kidney related health problems. The former is defined as enduring kidney damage or decreased kidney function, the latter as a sudden fall in the kidneys' excretion function. These two conditions are now recognised to be strongly associated. In this study, we investigate the influence of AKI on long-term kidney function amongst patients already diagnosed with CKD. We develop a longitudinal statistical model with Matern correlation structure and multivariate t-distributed stochastic components. We also specify three change-points that describe the typical trajectory of an AKI event. We use maximum likelihood estimation implemented with an expectation-maximisation algorithm to estimate the model parameters, and best linear unbiased prediction to predict the random effects. Our case-study uses data from an on-going cohort study of chronic kidney patients at Salford Royal NHS Foundation Trust, Salford, UK.

8.1 Contributed – Medical – Modelling case studies

Wednesday 9 September 3.15pm – 4.15pm

Predictive statistical ecological modelling of global tuberculosis burden

*Markus Elze, Elizabeth Allen, Katherine Fielding
London School of Hygiene and Tropical Medicine*

Tuberculosis (TB) is a global problem with approximately 9 million new cases and 1.5 million deaths each year. Halting and reversing TB incidence is part of the 2015 UN Millennium Development Goals (MDG). Reliable country-level estimates of TB incidence, prevalence and mortality are needed from 1990 to 2015 to assess the fulfilment of this goal. While TB surveillance today has vastly improved, reliable data on incidence, prevalence, and mortality are not available for all countries and all years since 1990.

As part of the MDGs, the WHO Global Task Force on TB Impact Measurement was created to conduct a review of methods for the global estimation of TB burden. So far, most published methods focused on deterministic adjustments to reported numbers, often relying on expert opinion. Here, we present work in progress on predictive statistical models using ecological data as an alternative to current approaches.

Mixed models incorporating country-level variables, such as wealth, health system status, education, and HIV burden, are used to predict TB incidence, prevalence and mortality. Options for random effects structures are considered. Results from our models are compared to current methodologies. The bounding of predictions and missing data are highly relevant issues in this setting and possible solutions are proposed. The benefits and drawbacks of the proposed models compared to both current methodologies and alternative statistical approaches, such as Bayesian models, are discussed.

8.1 Contributed – Medical – Modelling case studies

Wednesday 9 September 3.15pm – 4.15pm

Using Cox modelling to allocate donor livers to patients on the transplant list

Rhiannon Taylor, Elisa Allen, Alex Gimson, Dave Collett
NHS Blood and Transplant

More patients await a liver transplant in the UK than there are livers available for transplantation. This discrepancy requires donor livers to be allocated to patients in a way that makes the best use of such a scarce resource. Livers are currently allocated on a rota basis with transplant centres deciding which patient receives the liver based on clinical need. However, there is inequity in access to liver transplantation as sick patients may not be offered the liver due to the position of their transplant centre on the rota.

We used observational data from the UK Transplant Registry along with survival analysis techniques such as Cox proportional hazards regression modelling to investigate three alternative national allocation schemes, namely:

- allocation by *need* (patient prioritisation by predicted time to death on the waiting list without a transplant);
- allocation by *utility* (prioritisation by predicted time of survival post-transplantation);
- allocation by *transplant benefit* (predicted utility minus predicted need).

Livers are currently allocated to patients who are sicker, so censoring at time of transplant is informative in a time to death on the list analysis. Thus, our work dealt with informative censoring, missing-by-design values requiring multiple imputation techniques, and the implementation of statistical simulations to determine the impact of the proposed schemes.

The results of the simulations showed that the *transplant benefit* allocation scheme has the potential to save the lives of more than 40 patients awaiting a liver transplant per year who sadly die whilst waiting for a suitable liver.

8.2 Contributed – Official – Prices

Wednesday 9 September 3.15pm – 4.15pm

Investigation into the use of web scraped data to calculate consumer price indices

Chris Payne

Office for National Statistics, UK

UK consumer price statistics are compiled using approximately 180,000 price quotes collected once a month. Around 100,000 of these are collected in local shops by price collectors, whilst the remainder are collected via websites, emails and telephone calls. These price quotes represent around 700 items.

The Office for National Statistics is researching alternative data sources for its consumer price statistics. This includes using tools to scrape prices from retailer's websites. Since April 2014, ONS has scraped daily prices for 35 grocery items, collecting over 1.5 million prices quotes in 10 months. These types of datasets have the potential to improve the temporal coverage and product representation in consumer price indices.

Whilst web scraped datasets offer many benefits, significant challenges need to be overcome before they can be used in live production:

- Choice of compilation method for homogeneous prices.
- Matching of prices over time is problematic due to frequent website updates.
- Traditional index number formulae may not be appropriate for large datasets.
- The size of the data present handling and storage issues.

We describe early attempts to address these issues, and construct experimental indices for a selection of grocery items. Traditional index number formulae are used, such as the Jevons formula, (the geometric mean of price relatives), and the more recent GEKS methodology (the geometric mean of all bilateral Jevons indices). The research considers the effects of using all the data, mimicking a traditional collection, and sampling from the data. Indices of differing frequencies are also produced.

8.2 Contributed – Official – Prices

Wednesday 9 September 3.15pm – 4.15pm

How volatile are Fruit and Vegetable prices in the UK?

Nicky Pearce

Office for National Statistics, UK

European regulations (in particular 7001/2006) for the Harmonised Index of Consumer Prices (HICP) (the UK's CPI) require prices for products which show "sharp and irregular price changes" to be collected over more than one working week. The regulation specifically states that this applies to fresh fruit and vegetables. To comply with this regulation, the Office for National Statistics (ONS) launched a pilot study in 2010 that trialled an additional price collection for fruit and vegetables. This additional collection is being rolled out in phases across all 141 locations visited for CPI price collection. However, there is a question about just how volatile fruit and vegetable prices are in the UK. Work to investigate this used monthly CPI price quote data from 2013. Volatility was defined as a combination of frequency of price change and magnitude of price change. These two characteristics were examined separately for all food items. Following this, a series of one-way ANOVAs were carried out to ascertain whether fruit and vegetable prices are significantly more volatile than other food prices.

The results suggest that at an aggregated level, there is little evidence to suggest that fruit and vegetable prices show sharp and irregular price changes. At a lower level, there are a small number of particularly volatile fruit and vegetable items; however, not all of these are fresh, for example, premium potato crisps. Comparisons with other EU member states' HICP indices suggest that fruit and vegetable prices in the UK are among the least volatile in Europe.

8.2 Contributed – Official – Prices

Wednesday 9 September 3.15pm – 4.15pm

Location, location, location: New methodology for forming consumer price index location boundaries

*Liz Metcalfe, Donna Viney
Office for National Statistics, UK*

The UK's headline rate of inflation, the Consumer Price Index (CPI) is estimated using the prices collected for a shopping basket containing goods and services bought by households. Prices are collected monthly from retail outlets from a sample of 141 locations from across the UK.

Locations are intended to be representative of retail activity within the UK. The existing sample frame of locations was created in 2000 using a commercial database of 1,200 shopping centres, and partly updated in 2004. New outlets or outlets that have moved since 2004 are potentially no longer within the current location boundaries, and hence over time the sample frame may be becoming less representative of the UK's retail activity. The current locations cover around 76 per cent of total retail activity in the UK.

A new method of forming CPI Location Boundaries is presented here which aims to maximise retail activity and geographical location coverage within the UK. This method uses hotspot analysis to define location starting points, and then a location-allocation method combined with the formation of polygons to grow and define the location boundaries. The location boundaries are then 'clipped' back to follow real world features such as roads and railways.

Under this new method, the new sample frame covers over 84 per cent of the UK's retail turnover, and in particular has improved coverage in rural areas. A live pilot in 2016 will be used to evaluate this new methodology in practice.

8.3 Contributed – Environment – Ecology

Wednesday 9 September 3.15pm – 4.15pm

Parameter Redundancy in Ecological Integrated Population Models

Diana Cole, Rachel McCrea
University of Kent

To be able to fit or examine a parametric model successfully all the parameters need to be identifiable. If the parameters are non-identifiable the model can be rewritten in terms of a smaller set of parameters, and is termed parameter redundant. Parameter redundancy is not always obvious, in which case the definitive method for detecting parameter redundancy is a symbolic method. This involves calculating the rank of a matrix, which is expressed symbolically (see for example Cole et al., 2010, *Mathematical Biosciences*, 228, 16-30).

Many studies of ecological populations consist of more than one type of data. Such data sets can be examined simultaneously using an integrated population model. However combining several data sets does not necessarily mean all the parameters can be estimated. We present methods for determining which parameters can be estimated in integrated population models.

Models used in ecology are becoming more realistic but at the same time more complex. This poses a problem for the symbolic method as computers run out of memory calculating the rank of the appropriate matrix. Combining more than one model in an integrated population model adds to this complexity. We present extensions to the symbolic method that simplify the calculation considerably. The methods are illustrated using an integrated population model for four data sets on Common Guillemots. This involves capture-recapture data, mark-resight-recovery data, productivity data and count data.

8.3 Contributed – Environment – Ecology

Wednesday 9 September 3.15pm – 4.15pm

Do rainforests contain communities? Characterising dependence structures in massive populations of point processes.

James Martin, David Murrell, Sofia Olhede
University College London

Characterising biodiversity is an important problem in statistical forest ecology as evidence suggests that forests containing more species will be more abundant as a whole, and therefore more useful in drawing CO₂ from the atmosphere. In a given location, it is therefore important to understand which species can survive and propagate, and which species cannot. Central to the understanding of biodiversity is the question of how individuals interact with each other, both within the same species and between different species, and much of the literature to date has maintained a species-level approach to analysing biodiversity, describing the interaction between all pairs of species in a given forest population. Given a population comprising many species, it is relevant to ask whether a community structure exists: are there groups of species that successfully cohabit, and can these groups be reliably identified? Through addressing this question, we provide a more concise, population-level description of biodiversity.

Multivariate spatial point process models provide a flexible framework for analysing multi-species forest populations, as existing tools for characterising the multivariate second-order dependence structure can be used to quantify within- and between-species interaction. We assume a spatial Cox process model, and we relax the common, yet unrealistic assumption of rotational homogeneity, or isotropy, in the marginal dependence structures. Once we establish this framework, we construct a method for discovering communities within massive populations of point processes. This work is demonstrated on rainforest data from the Barro Colorado Island in Panama, comprising approximately 275,000 individuals from over 300 species.

8.3 Contributed – Environment – Ecology

Wednesday 9 September 3.15pm – 4.15pm

Detecting heterogeneity in capture-recapture models: diagnostic goodness-of-fit tests and score tests

Anita Jeyam, Rachel McCrea, Diana Cole, Byron Morgan, Roger Pradel
University of Kent

Understanding the dynamics of animal populations is a key component in interpreting the effects of environmental change. The strategic importance of such research is evident from current government policies.

Capture-mark-recapture (CMR) is a technique used to study a species: animals are uniquely marked and released, then records are made of whether individuals are recaptured at subsequent sampling (capture) occasions; the resulting data are known as CMR data.

Capture-recapture models have increased in complexity in order to be more realistic and provide more accurate biological interpretation.

We focus on the case where the data exhibit heterogeneity in capture (some animals are captured very often, others rarely), which is modelled using mixture models (Pledger et al, 2003; *Biometrics*, **59**:786-794).

We investigate the use of goodness-of-fit tests to find out whether this type of heterogeneity can be detected and specifically pinpointed: for instance, does heterogeneity in capture affect the tests differently from a combination of trap-dependence (capture at a specific occasion affects the probability of being captured at the subsequent occasion) and transience (presence of animals just passing through the sampling site)?

We use simulations and real applications to assess the performance of the approach proposed in Péron et al. (2010; *Oikos*, **119**:524-532). Based on these simulations, we then investigate the same questions, through the use of score-tests, looking at model selection in conjunction with goodness-of-fit testing, extending the approach of McCrea et al. (2015, in preparation). Lastly, we compare the performance of score tests and goodness-of-fit tests in this framework.

8.4 Contributed – General – Match fixing, Betting and Forecasting

Wednesday 9 September 3.15pm – 4.15pm

Smashing the racket: Detecting match-fixing in tennis via in-play betting irregularities

Tim Paulden
ATASS Sports

An unfortunate consequence of the recent growth in tennis betting markets has been a heightened incentive for match-fixing - particularly at lower levels of the sport, where earnings are modest, and where the market for one match may dwarf the prize money for an entire tournament. Against this backdrop, an unscrupulous player may be tempted to profit from their position of influence by secretly agreeing to "throw" a particular match, with a complicit third party betting on the result.

The ability to detect - and help eliminate - match-fixing of this kind is valuable from both a moral and commercial standpoint, and the fledgling field of "forensic sports analytics" (Rodenberg 2014; Reade & Akie 2013) seeks to identify such corruption by using statistical modelling to highlight anomalous odds movements. However, existing work has focused exclusively on pre-match markets, rather than the evolution of in-play probabilities.

This talk summarises the results of a match-fixing study spanning more than 6,000 in-play tennis betting markets. After developing a natural point-by-point probabilistic model, featuring novel mechanisms for selecting parameters robustly from the data, we demonstrate that the observed market trajectories correlate extremely closely with this model. We argue that substantial discrepancies represent a "red flag" that something is amiss - either an injury, or something more covert. We also highlight the considerable challenges of accurately reconciling non-timestamped point-by-point data with the market probabilities. Finally, we provide visualisations of recent matches where the market evolved pathologically, and assess the evidence that these matches were fixed.

8.4 Contributed – General – Match fixing, Betting and Forecasting

Wednesday 9 September 3.15pm – 4.15pm

What can gambling machine data tell us about betting behaviour?

David Excell, Heather Wardle
Featurespace Ltd

Bookmakers' betting machines continually attract political and media attention over their impact on problem gambling. Featurespace and NatCen partnered with the Responsible Gambling Trust to investigate harmful patterns of play on gaming machines, and draw implications for intervention.

We'll present the methods and results of this ground-breaking investigation, linking industry-held data from the five largest UK bookmakers with surveys of loyalty card customers (which measured customers' Problem Gambling Severity Index score as a proxy for harmful play), integrating research methods for 10 billion individual gaming machine events. The data set included 6.7 billion bets and 333,000 customers.

We harnessed this huge data set to model actual gaming play, measure theoretical markers of harm (e.g. faster gaming), survey loyalty card customers (matching 4,001 responses with transactional data) and explore consumer interventions.

We produced two predictive models, exploring the statistical relationships between the data and the customer surveys. The results showed it is possible to distinguish between problem gamblers and non-problem gamblers in industry data:

- **Player model:** behavioural analyses in loyalty card holder data - 66% improvement over the current baseline model.
- **Session model:** proxy measurements for anonymous players rather than individual players - 550% improvement in accuracy of detecting problem gamblers over the industry standard.

The research demonstrated that a combination of variables are needed to identify problem gamblers, in contrast to proposed policy suggestions of regulating individual parameters (e.g. stake size).

We'll assess approach limitations, including data skewedness, and explore the challenges of incorporating big data into social scientific investigations.

8.4 Contributed – General – Match fixing, Betting and Forecasting

Wednesday 9 September 3.15pm – 4.15pm

Forecasting Incoming Volume into a Call Centre

Andrew Zelin

Bupa

Call Centres represent a significant growth area of the world's economy and for most of the largest organizations, are used as the mainstay for handling the majority of their calls coming in. Therefore, the need for organizations to be able to accurately forecast how much work they anticipate coming in is becoming increasingly critical, so that the right number of staff can be at their desks at the right times to take the calls. The need to schedule and plan capacity on the back of these forecasts is essential as well as run a fine-line between optimizing the customer experience and keeping running costs low.

This presentation describes how it has been possible to obtain daily estimates of inbound call volume to within 5% of the actual volumes by looking at historical data, presenting this as a step-by-step guide that operational planners can take when they need to achieve this. Four years of daily volume data and recognized statistical techniques (including tests of model adequacy) have been used to assess what is typical for a Monday, or Tuesday, what is typical for weekends and bank holidays, to look at month and seasonal effects and through tracking the number of individual callers, possible "bounce-backs" on days following closed days or holidays or very busy periods. Also assessed are the variations at different times of the year / in school holidays and within-month as well as what happens with one-off changes in policy, the "customer base" or marketing campaigns.

8.5 Contributed – Methods & Theory – Modelling Complex systems

Wednesday 9 September 3.15pm – 4.15pm

Evaluating the performance of ensemble forecasts

Christopher Ferro, Stefan Siegert
University of Exeter

An ensemble forecast is a collection of point forecasts, often interpreted as a simple random sample from an unknown probability distribution. In many applications, we would like these underlying distributions to be well calibrated (so that predictands also behave like random draws from the distributions) and sharp (so that the distributions express low uncertainty about the predictands). If the underlying distributions were known then we could assess their calibration and sharpness using tools, such as proper scoring rules and reliability diagrams, that have been developed for evaluating probability forecasts. We show that naively applying these tools to ensemble forecasts fails to reward ensembles whose underlying distributions are well calibrated and sharp. We introduce modified scoring rules and reliability diagrams that overcome this deficiency, and we illustrate their use by evaluating the performance of ensemble weather forecasts.

8.5 Contributed – Methods & Theory – Modelling Complex systems

Wednesday 9 September 3.15pm – 4.15pm

Generalised additive and index models with shape constraints

Yining Chen, Richard Samworth
University of Cambridge

We study generalised additive models, with shape restrictions (e.g. monotonicity, convexity, concavity) imposed on each component of the additive prediction function. We show that this framework facilitates a nonparametric estimator of each additive component, obtained by maximising the likelihood. The procedure is free of tuning parameters and under mild conditions is proved to be uniformly consistent on compact intervals. More generally, our methodology can be applied to generalised additive index models. Here again, the procedure can be justified on theoretical grounds and, like the original algorithm, possesses highly competitive finite-sample performance. Practical utility is illustrated through the use of these methods in the analysis of real data on decathlon. Our algorithms are publicly available in the R package "scar", short for Shape-Constrained Additive Regression.

8.5 Contributed – Methods & Theory – Modelling Complex systems

Wednesday 9 September 3.15pm – 4.15pm

Bayes linear analysis for Bayesian optimal experimental design

*Matthew Jones, Michael Goldstein, Philip Jonathan, David Randell
Durham University*

In many areas of science, models are used to describe attributes of complex systems. These models are generally highly complex functions of their inputs, and can be computationally expensive to evaluate. Often, these models have parameters which must be estimated using data from the real system. We address the problem of using prior information supplied by the model, in conjunction with prior beliefs about its parameters, to design the collection of data such that it is optimal for decisions which must be made using posterior beliefs about the model parameters. Optimal design calculations do not generally have a closed form solution, so we propose a Bayes linear analysis to find an approximately optimal design; we motivate our approach by considering the optimal specification of measurement locations for remote sensing of airborne species. We find that the analysis is effective in ruling out poor designs and finding parts of the design space which give relatively low risks.

8.7 Contributed – Data Science – Elections

Wednesday 9 September 3.15pm – 4.15pm

Predicting the Results of the United Kingdom General Election: analysis of 2010 and 2015

Timothy Martyn Hill
Liverpool Victoria

Metrics are created by others and used to predict the outcome of the UK General Election. Which ones are the best? To answer this question we list the common metrics (exit polls, opinion polls, betting odds, betting spreads, party funding, campaign spending, academic models, bookmaker models...). We note the different definitions of the final result. We discuss methods of comparing the metrics to the final result. We then use the chosen method to compare those metrics to the final results, both within classes (which poll the best, which model the best model..), between classes (which is better: poll, model, odds, funding, models...), and over time (which is better the day before, the month before, six months before...) Finally we present the conclusions.

8.7 Contributed – Data Science – Elections

Wednesday 9 September 3.15pm – 4.15pm

A Dynamic, Hierarchical, Bayesian approach to Forecasting the 2014 US Senate Elections

Roberto Cerina, Gianluca Baio
University College London

In this paper I will show that the 2014 Senate election was uncertain up to the very end, and Republicans' fortunes turned in the last few weeks. A bias in the polls favoured the Democratic party and led to unexpected uncertainty over the results. This uncertainty was gone by the final week of polling and a Republican takeover of the Senate was close to unquestionable. The extent to which the Republican wave demonstrated itself was, however, unexpected.

In 2012, Drew A. Linzer presented a dynamic hierarchical Bayesian state level forecasting model to predict the US Presidential election. Here I present an adaptation of Linzer's model. My model's peculiarity is that it focusses on state level structural fundamentals for its historical forecast, as opposed to national factors, and adopts a full-Bayesian specification to propagate the uncertainty that exists on these structural variables onto my final predictions. The model combines historical trends with increasingly precise polls.

At the end of every week during the election campaign, I take a summary of published polls and fit the model to calculate the distribution of voter preferences for every week in the campaign. Initially, the predictions are as good as the historical forecast, but they get closer to the actual election results as more and more polls are encoded. I concentrate on the Republican vote share, so that I use a hierarchical specification to detect national swings common to conservatives. Simulation techniques are then used to calculate the probability that Republicans take over the Senate.

8.7 Contributed – Data Science – Elections

Wednesday 9 September 3.15pm – 4.15pm

Checking the UK General Election fact checkers

Andrew Garratt

"The next election in May 2015 will see a growth of fact-checking that will be 'coming out of our ears', according to a roundtable at Nesta on the use of evidence for the General Election."

NESTA - <http://www.nesta.org.uk/blog/zombie-facts-and-use-evidence-2015-general-election>

Political claims were fact checked by several organisations in the 2015 UK General Election.

The output of fact checkers was largely followed using Twitter, and by following their web sites.

This presentation reviews how the organisations fared in areas including: the numbers and types of facts and claims checked; the extent to which sources were used and referenced; what explanations were given of choices of baselines and methods of calculations; and contextual discussion.

In particular, the presentation will examine the errors made by the fact checkers themselves, their types and impact, and the ways in which these were acknowledged, corrected and explained.

A comparison of those types of error in political claims and those made by fact checkers will be made to identify those issues which are common to both (both statistical and logistical) and those issues which are particular to one or other group. Comments will be made as to how these are addressed through the period to the next General Election.

9.1 Invited – Medical – Out of Practice Primary Care

Wednesday 9 September 4.40pm – 6pm

Statistician behind Bars: by design and on trial

Sheila Bird

MRC Biostatistics Unit, Cambridge

I shall describe how a quarter century of surveillance designs (with associated biological sample), record-linkage studies, and bespoke "questionnaires" - a phrase coined from Hill and Doll - have improved prisoners' access to harm reduction (eg Hepatitis B immunization), contributed to changed policy in prisons (eg put an end to random mandatory drugs testing), got us barred, yet quantified a 7 times higher risk of overdose death soon after prison-release, and eventually enabled three musketeers to mount the pilot N-ALIVE Trial in England to test whether those randomized to receive naloxone-on-release have 30% fewer opioid-related deaths in the 4-weeks post-release than controls (prior estimate: 1 in 200).

Even before the N-ALIVE Trial's first randomization, however, Scotland became the first country to introduce take-home-naloxone as a funded public health policy. Wales followed in May 2011. I shall describe the trials and tribulations of Scotland's closely-monitored evaluation of its 2011-15 national naloxone policy, which is complicated because Scotland's policy was introduced against a still-rising trajectory of age-related opioid-deaths.

Is it impact for an RCT to be overtaken by the policy it seeks to inform?

9.1 Invited – Medical – Out of Practice Primary Care

Wednesday 9 September 4.40pm – 6pm

Use of a stepped-wedge design to compare prescribing errors in community pharmacies using the NHS electronic prescription service compared to paper prescriptions

Matthew Boyd, Sarah Armstrong, Rajnikant Mehta
University of Nottingham

Objectives

To compare prevalence and types of dispensing errors and pharmacists' labelling enhancements, for prescriptions transmitted using the NHS electronic prescription service (EPS) versus paper prescriptions in community pharmacy.

Method

Dispensed prescription medicines were reviewed by a qualified researcher (pharmacist technician) across 15 English pharmacies and discrepancies between the items requested and those dispensed were recorded on a standardised proforma noting prescription type (electronic/paper based). The study used a stepped wedge design adapted to the natural (rather than research-led) deployment of the intervention.

Clinical Results

16357 dispensed items were reviewed. Labelling and content errors were identified in 5.4% and 1.4% of items respectively; 13.6% had positive enhancements. Items sent via the EPS had a higher prevalence of labelling errors than paper prescriptions, but a similar rate of content errors and enhancements. The increase in labelling errors was mainly accounted for by one pharmacy omitting the indication from the label and a sensitivity analysis excluding these cases revealed no remaining difference.

Conclusions

Items transmitted electronically had higher prevalence of labelling errors, but this was predominantly due to local practice in a single pharmacy. Approximately one in seven prescription items were enhanced by community pharmacists. Medicines supply professionals should work together to agree how items should be dispensed and labelled to best reap the benefits of electronically transmitted prescriptions.

Our experiences in using a stepped wedge in the context of the natural deployment of an intervention are discussed and practical advice for other research studies considering this approach is offered.

9.1 Invited – Medical – Out of Practice Primary Care

Wednesday 9 September 4.40pm – 6pm

We only randomise to treatment here! Is the “Cohort Multiple Randomised Controlled Trial” a better alternative to the pragmatic RCT?

*David Reeves, Mark Hann, Peter Bower
University of Manchester*

The “cohort multiple randomised controlled trial” (cmRCT) design is a fairly recent innovation that aims to overcome some of the limitations of traditional pragmatic RCTs. The design is based upon first recruiting a large cohort of patients who are followed up longitudinally at regular time-periods, then randomly selecting smaller subsets of these to be offered trial interventions. Outcomes in this subset after the intervention are then compared with outcomes in the remainder of the cohort, who constitute the control group. With a large cohort, multiple RCTs can be conducted using different random draws of patient subsets.

At Manchester we are using a cmRCT design as a means of creating a pool of participants for testing out a number of public health initiatives. We have recruited a cohort of 4,000 patients and are in the process of recruiting 250 of these into an embedded trial, “CLASSIC PROTECTS”, of a centrally provided telephone-based health coaching scheme supported by a multidisciplinary health and social care team.

In the process, we have found that as well as having a number of appealing features the cmRCT design also raises a number of interesting and important methodological issues. In this talk we will describe our experience of the cmRCT and discuss these issues, including (i) power and sample size relative to a comparable pragmatic RCT; (ii) the impact of non-consent and non-adherence; (iii) the treatment effect being estimated and implications for analysis; (iv) other potential sources of bias.

9.3 Invited – Environment – Recent advances in Statistical Ecology

Wednesday 9 September 4.40pm – 6pm

Integrated population model selection in ecology

*P.T. Besbeas, R.S. McCrea, B.J.T. Morgan
University of Kent*

In ecology it is often the case that aspects of the demography of wild animals are studied using different data collection methods, resulting in multiple data sets. Typically, component models share some common demographic probabilities and rates. Integrated population modelling occurs when likelihoods arising from independent data sets are multiplied together to form a single, joint likelihood; see Besbeas et al (2002).

The approach is now widely used, but there has been no formal assessment of how model selection should be addressed.

This talk compares alternative methods for model selection in integrated population modelling using classical inference, considering component likelihoods separately and in combination. I will focus on the important case of determining the age-structure for annual survival probabilities when two component data sets are involved. The work is motivated by real data sets, and evaluated by simulation. A senescence case study is also provided.

I will compare the use of procedures based on likelihood-ratio tests as well as AIC and will demonstrate that it is generally better to do model selection using likelihood-ratio tests and the joint likelihood, rather than use a component likelihood.

9.3 Invited – Environment – Recent advances in Statistical Ecology

Wednesday 9 September 4.40pm – 6pm

Bayesian non-parametric population ecology models

Eleni Matechou, François Caron
University of Kent

We present a Bayesian nonparametric approach for modelling capture-recapture data for open non-resident populations. Arrival times of individuals are assumed to be drawn from a Poisson process with unknown intensity function, which is modelled via a flexible nonparametric mixture model. The proposed capture-recapture model can be used to estimate the population size and the times of arrival and departure of all individuals in the population, caught or not. We apply the model to two data sets: a data set of breeding great crested newts (*Triturus cristatus*), demonstrating the staggered arrival of individuals at the breeding ponds, and a data set of migrating reed warblers (*Acrocephalus scirpaceus*), demonstrating the arrival of migrating flocks of individuals at the stopover site.

9.3 Invited – Environment – Recent advances in Statistical Ecology

Wednesday 9 September 4.40pm – 6pm

Ecological applications of hidden Markov and related doubly stochastic processes

Roland Langrock

University of St Andrews

Doubly stochastic processes involve an observed stochastic process which in some way is driven by a second, unobserved stochastic process, usually a Markov process. Corresponding models, such as hidden Markov models, general state-space models, Markov-switching regression models or Markov-modulated Poisson processes, have many applications in ecology, where the unobserved process could correspond, for example, to the behavioural state of an animal, to an animal's condition or to the not directly observable size of an animal population. In this talk, I will first demonstrate how the relatively simple yet very powerful hidden Markov model machinery can be exploited to make inference tractable in many different types of doubly stochastic processes. I will then demonstrate the usefulness of these methods in ecology by sketching several areas of application, including animal movement modelling, animal abundance estimation and capture-recapture studies.

9.5 Invited – Methods & Theory – Copula Modelling and Applications

Wednesday 9 September 4.40pm – 6pm

Using local dependence to evaluate goodness of fit in copula regression models

Mario Cortina Borja, Eirini Koutoumanou, Angie Wade
UCL Institute of Child Health

Copula regression models are increasingly being used in applied statistics to study continuous bivariate outcome variables; these models allow us to specify linear predictors for the parameters of the marginal distributions as well as for the copula dependence parameter. However there are few diagnostics tools used to assess goodness of fit in this class of models. In particular, if the model were correctly specified then normalised residuals should appear as uncorrelated standard bivariate random variables. In this talk we discuss using local dependence maps (Jones and Koch (2003) *Statistics and Computing*) and methods for local Gaussian correlation (Berentsen, Kleppe and Tjøstheim (2014) *J of Statistical Software*) to evaluate departures from the model assumptions. We illustrate this methodology applying it to modelling epidemiological and clinical data.

9.5 Invited – Methods & Theory – Copula Modelling and Applications

Wednesday 9 September 4.40pm – 6pm

Bayesian Copula Modelling in the Presence of Covariates

*Julian Stander, Luciana Dalla Valle, Charlotte Taglioni, Mario Cortina Borja
University of Plymouth*

Copula models separate the dependence structure in a multivariate distribution from its univariate marginals, so overcoming many of the issues associated with commonly used statistical modelling methods by allowing, for example, different complex asymmetric dependencies and tail behaviours to be modelled. We discuss the modelling of bivariate data using copulas, of which there are now a rich choice. The parameter or parameters of the copula density are modelled as a function of a covariate using a natural cubic spline. Working in the Bayesian framework, we perform inference on the natural cubic spline and an associated smoothing parameter. We also discuss the choice of the copula density itself. We illustrate our approach using data from finance and medicine. We outline the extension of our methodology to more than one covariate and to multivariate data.

9.5 Invited – Methods & Theory – Copula Modelling and Applications

Wednesday 9 September 4.40pm – 6pm

Copula mixed effect models for multivariate meta-analysis of diagnostic test accuracy studies

Aristidis K. Nikoloulopoulos
University of East Anglia

A generalized linear mixed model (GLMM) is currently recommended to synthesize diagnostic test accuracy studies. We propose copula mixed models for multivariate meta-analysis in this context. Our general model includes the GLMM as a special case and can also operate on the original scale of sensitivity and specificity. Summary receiver operating characteristic curves are deduced for the proposed model through quantile regression techniques and different characterizations of the bivariate random effects distribution. Our general methodology is demonstrated with an extensive simulation study and illustrated by re-analysing the data of several published meta-analyses. Our study suggests that there can be an improvement on GLMM in fit to data and makes the argument for moving to copula random effects models.

References

Nikoloulopoulos, A. K. (2015a) A mixed effect model for bivariate meta-analysis of diagnostic test accuracy studies using a copula representation of the random effects distribution. *Statistics in Medicine* (in press). DOI: 10.1002/sim.6595.

Nikoloulopoulos, A. K. (2015b) A vine copula mixed effect model for trivariate meta-analysis of diagnostic test accuracy studies accounting for disease prevalence. *Statistical Methods in Medical Research* (in press). DOI: 10.1177/096228021559676910.

Nikoloulopoulos, A. K. (2015c) CopulaREMADA: Copula mixed effect models for bivariate and trivariate meta-analysis of diagnostic test accuracy studies. Available from: <http://CRAN.R-project.org/package=CopulaREMADA>, R package version 0.9.

9.6 Invited – Data Science – Big Data: the challenges in using large non-random data sets

Wednesday 9 September 4.40pm – 6pm

Big data, big changes for official statistics?

Jane Naylor

Office for National Statistics, UK

The Government Statistical Service (GSS) is responsible for collecting and publishing statistics related to the economy, population and society at national, regional and local levels. These statistics are used within central and local government to underpin policy making and to plan services. The GSS recognises the importance of examining the potential of big data sources and related technologies. This presentation will cover progress being made to investigate the potential advantages of using big data, to understand the challenges with using these data sources and to establish a longer term strategy for big data within official statistics. Particular focus will be given to case studies from the GSS such as

- Web scraping data for price statistics
- Analysis of smart-type meter data
- Using geo-located Twitter data to provide intelligence on migration

Each case study will illustrate the potential benefits that big data could bring to the production of official statistics as well as highlighting the challenges, in particular the methodological and quality issues and how these can be managed if big data is going to make a big impact on official statistics.

9.6 Invited – Data Science – Big Data: the challenges in using large non-random data sets

Wednesday 9 September 4.40pm – 6pm

Challenges and Perspectives in Exploring Complex & Big Data

Maria Fasli
University of Essex

With the rapid growth of computational devices being used in almost all aspects of human activity, we have seen an explosion in the volumes of data, both structured and unstructured, being generated in a range of domains. This talk will be describing some of the issues and challenges arising in analysing and modelling big and complex data from a computer science perspective. The presentation will discuss work that has been using artificial intelligence techniques such as agent-based and multi-agent systems, to explore and understand complex and big data in different domains of application, ranging from understanding the dynamics of electronic markets to individual user behaviour.

9.6 Invited – Data Science – Big Data: the challenges in using large non-random data sets

Wednesday 9 September 4.40pm – 6pm

Big data is more than behaviour: why business needs to get smarter about extracting insights from big data

Colin Strong

Big data promises much for companies seeking to gain competitive advantage. McKinsey, for example, estimate that retailers using big data can increase their margin by more than 60%. No wonder that management consultancy AT Kearney forecast the value of the big data tech market will be \$114B by 2018. However, we are also starting to see some disillusionment with the big data project. Gartner estimate that through 2017, 60% of big data projects will fail to go beyond piloting and experimentation and will be abandoned. This is often used as an argument for greater spend on technology to ensure that data is pulled together into useable formats.

However, whilst data aggregation and preparation is a necessary condition for generating value, it is questionable whether this is in itself sufficient. As Nate Silver points out that 'The numbers have no way of speaking for themselves'. The challenge, therefore, is what inferences we can justifiably draw about consumers from their behavioural data. Can we merely describe their behaviours, or can we start to infer explanations about the causal factors driving their behaviours. And can we make wider interpretations concerning consumers' personal, social and cultural lives? This review of the literature and practitioner experience draws together philosophy of knowledge with the practical concerns of businesses seeking to make use of big data.

The paper reaches the conclusion that current practices are limited and outlines a model for business that highlights the opportunities for maximising their big data investments

Special Session – Learning from Multivariate Data: A tribute to Wojtek Krzanowski
Wednesday 9 September 4.40pm – 6pm

Summarising classifier ROC curves

David Hand
Imperial College London

Evaluation lies at the core of supervised classification. Comparison between methods, selection of methods, parameter estimation, and feature choice all implicitly require evaluation. Many different classifier performance summary statistics have been used, but one is particularly popular - the area under the ROC curve, or its equivalents the Gini index and the Mann-Whitney-Wilcoxon statistic. Unfortunately, this has a major weakness when used for evaluating classifier performance, which means it may be unfit for purpose.

Special Session – Learning from Multivariate Data: A tribute to Wojtek Krzanowski
Wednesday 9 September 4.40pm – 6pm

Multivariate methods and ecology

Byron Morgan
University of Kent

Ecological data are necessarily multivariate. In this talk a particular connection is described, involving N-mixture models (Royle, 2004). These have become a popular way of estimating abundance using count data, obtained from repeated visits to multiple sites. An attraction is the low cost and effort required for the data collection, and use with citizen-science data. Recent research (Dennis et al, 2015) has shown that the models are equivalent to multivariate discrete distributions. A useful computational consequence of this is that it avoids the arbitrary truncation of summations needed for the N-mixture formulation. Additionally a diagnostic test can be devised which indicates when the model may result in misleading results. Originally designed for closed populations, the model extends also to open populations (Dail and Madsen, 2011), and may be fitted efficiently using hidden Markov methods. Avenues for future research are outlined.

References

Dail, D. and Madsen, L. (2011) Models for estimating abundance from repeated counts of an open metapopulation, *Biometrics*, **67**, 577-587.

Dennis, E.B., Morgan, B.J.T. and Ridout, M.S. (2015) Computational aspects of N-mixture models. *Biometrics*, **71**, 237-246.

Royle, J.A. (2004) N-mixture models for estimating population size from spatially replicated counts. *Biometrics*, **60**, 108-115.

10.2 Contributed – Official – Official Statistics Methods

Thursday 10 September 9.10am – 10.10am

Behind the Marmot curve: Why and how do some communities do better or worse than their levels of disadvantage would predict?

*David Buck, David Maquire
King's Fund*

The report "Fair Society, Healthy Lives", led by Michael Marmot, drew together many different sources to create a compelling argument for a greater focus on health inequalities in and the wider determinants of health, such as education, income inequality and work in England. This has been influential with policy-makers and local authorities as responsibilities within public health change.

The iconic chart from this work is the so-called "Marmot Curve", which showed the relationship between life expectancy (LE), disability-free life expectancy and income deprivation across medium super-output areas in England. This chart and its implications have become synonymous with the review.

However, the data on which the chart is based, 1999-03 is old, and addressed only one variable (income deprivation) with regards to health. We have updated the original Marmot Curve, and placed it in a multivariate setting, bringing together a unique collection of data, using linear and probit regression models to understand the contributions of wider determinants, lifestyles and measures of public service quality to inequalities in LE. We also test the persistence of under and over-performing communities over time, and what determines "outlier" status: why some communities do better than others.

Several clear results emerge which suggest a focus on unemployment, older people's deprivation and behavioural changes to a healthier lifestyle are key in helping people in the most deprived areas live longer. Furthermore, we can present useful and important lessons when trying to analyse and bring together multiple deprivation and other data sources into a multiple regression model.

10.2 Contributed – Official – Official Statistics Methods

Thursday 10 September 9.10am – 10.10am

Handling attrition and non-response using weighting - what methods make a difference?

Matthew Greenaway

Office for National Statistics, UK

Estimating finite population characteristics from social surveys inevitably involves worrying about non-response and attrition bias. Standard practice at most survey organisations is to attempt to remove some of this bias by using 'model-assisted' estimation, whereby a weight is calculated for each case which accounts for both the selection probability and some auxiliary information. This auxiliary information can include sample characteristics ('sample-based weighting') and/or external information such as census-based population totals ('population-based weighting').

All ONS social surveys utilise population-based weighting, but only some utilise sample-based weighting. Recent work at ONS has investigated under what scenarios sample-based weighting makes an important difference. This involved using sample characteristics to model non-response and attrition processes separately for a number of surveys and using the resulting response propensities to adjust the survey weights. The impact on a wide variety of estimates, including on headline labour market statistics, wellbeing statistics, and health statistics, was then evaluated.

In general, it appears that if population-based weighting is used, adding in sample-based weighting as well makes little difference in most circumstances - the non-response or attrition mechanism is usually well-captured by variables such as age, gender and region, for which population-based weighting can adjust. Sample-based weighting can make a difference in longitudinal surveys where a specific attrition process can be identified which is directly related to outcome variables - for example, in health-related surveys, where an individual's health may influence their consent to a follow-up study.

10.2 Contributed – Official – Official Statistics Methods

Thursday 10 September 9.10am – 10.10am

A 50-year journey: measuring changes in personal travel using the National Travel Survey

*Delphine Robineau, Matthew Tranter, Glenn Goodman
Department for Transport, UK*

2015 marks the 50th anniversary of the first data collection of the National Travel Survey. Founded on a complex methodology, the NTS provides a unique source of evidence on how, why and when people travel in England. This talk will present the survey, its methodology, and explain what makes this data source such a unique and important evidence source for policy-making. The general objective of the talk will be to show how the picture of personal travel has changed in 50 years. We will analyse long term trends in trips, time and distance travelled. We will also discuss how our modes of travel have changed, particularly the evolution of car use, walking and cycling. Then, we will investigate the changes in our reasons for travelling. We will explore potential explanations for the evolution of travel behaviours in time - from key demographics (age, gender, income, type of residence) to new elements of lifestyle (homeworking, online shopping...). The talk will use descriptive statistics and infographics to visually engage with the audience. It will give the audience a broad understanding of long term trends in personal travel, propose possible explanations, and encourage the public to make their own use of the statistics and the data.

10.3 Contributed – Environment – Seabirds, fish and seedlings

Thursday 10 September 9.10am – 10.10am

Gaussian Processes in Species Distribution Modeling: Case Studies from the Baltic Sea

Jarno Vanhatalo, Meri Kallavuo, Lari Veneranta
University of Helsinki

The main objectives in species distribution modeling (SDM) are to predict the spatio-temporal occurrence or abundance pattern and to identify the environmental variables that best describe these patterns. This information can then be used to, e.g., predict the species distribution under changing environment and to aid environmental managers in spatial planning. Traditional examples of SDMs are generalized linear and additive models. Recently it has been proposed that SDMs could be built using Gaussian processes (GP). GP is a stochastic process that defines probability distribution over functions. They have received much interest in machine learning and statistics in recent years due to their non-parametric nature which allows flexible and versatile modeling. Linear and additive models can be seen as a special case of a GP. In this talk we present how GPs can be used in SDMs and what are the benefits and limitations of this approach. Our model combines a non-linear predictor with a spatio-temporal random effect under the common GP framework. We show how Bayesian analysis can be done efficiently using expectation propagation and Laplace algorithms. We demonstrate the method by analyzing and mapping the reproduction habitats of 3 commercially important fish species in the northern Baltic Sea. We also discuss how this information can be used to inform environmental managers in marine spatial planning.

10.3 Contributed – Environment – Seabirds, fish and seedlings

Thursday 10 September 9.10am – 10.10am

There is more to climate than the North Atlantic Oscillation: a new perspective from climate dynamics and climatology-specific methods to explain the variability in population growth rates of a long-lived seabird

Michel d. S. Mesquita, Kjell Einar Erikstad, Hanno Sandvik, Robert T. Barrett, Tone Kristin Reiertsen, Tycho Anker-Nilssen, Kevin I. Hodges, Jürgen Bader

Bjerknes Centre for Climate Research, Norway

Ecosystems worldwide are changing rapidly as a consequence of global warming, yet our understanding of the consequences of these changes on populations is limited. The North Atlantic Oscillation (NAO) has been used as a proxy for 'climate' in many ecological studies, but has not always explained the patterns of variation in populations examined. Other techniques to study the relationship between ecological time series and climate are therefore needed. A standard climatology method is to work with point maps, where point correlation, point regression or other techniques are used to identify regional hotspots that can explain the time series variability. In this study we have used climate-related techniques and analysed the yearly variation in the population growth of a Common Guillemot *Uria aalge* colony in the Barents Sea over 30 years. We show that while the NAO does not explain this variation, point analysis can help identify indices that can explain a significant part of it. These indices are related to changes of mean sea level atmospheric pressure in the Barents Sea via the Pacific - forming a teleconnection-type pattern. The dynamics are as follows: in years of an anomalous low-pressure system in the Barents Sea, the population growth rate increases. These low-pressure systems are a source of heat transport into the region and force upwelling mixing in the ocean, thus creating favourable conditions for a more successful survival and breeding of the Common Guillemot.

10.3 Contributed – Environment – Seabirds, fish and seedlings

Thursday 10 September 9.10am – 10.10am

Perturbed grid processes and plot sampling of seedling density

Aki Niemi

University of Jyväskylä

Objectives: Planted seedlings in a forest regeneration area are aimed to be distributed as regularly as possible, with their inter-tree distance corresponding to the targeted planting density. I will present perturbed grid processes as a new spatial point process model for seedlings. The model is applied in a sampling study where seedling density is estimated from circular plots in systematic design.

Methods and Models: The model is constructed from independent random perturbations of a regular grid. Mortality of planted seedlings and arrival of naturally regenerated seedlings are taken into account by random thinning and Poisson superposition, respectively.

Results and Conclusions: I will present pair correlation functions for various examples of the model. The sampling study results in expected error variances under each model, area of the seedling stand, and sampling design by plot size and distance between the plots.

10.4 Contributed – General

Thursday 10 September 9.10am – 10.10am

Semiparametric Least Squares Estimation of Single-Index Model with Nonstationary Regressors

Ryota Yabe

Hitotsubashi University

A nonlinear regression model with nonstationary regressors has attracted increasing attention, because macroeconomic empirical studies often needs to deals with nonlinear functions such as a convex function and concave one. Though the nonlinear model with nonstationary regressors is often modelled as parametric or nonparametric one, each model has a problem.

The parametric model needs prior knowledge about functional form that economic theory does not necessarily provide. In the nonparametric model, the number of the regressors must be restricted to be one or two by the technical reason to develop asymptotic theory. This talk proposes the single-index model that does not need prior knowledge about functional form and allows the arbitrary number of the regressors. Since this model includes various models such as linear cointegration, probit and logit ones, this model can be applied to various empirical studies.

This paper applies semiparametric least squares (SLS) estimation techniques proposed by Ichimura (1993) to the single-index model with the nonstationary regressors and derives consistency and asymptotic distribution of the SLS estimator. We also demonstrate its performance via simulation and empirical analysis by using Japanese data.

10.4 Contributed – General

Thursday 10 September 9.10am – 10.10am

Asymptotics for Estimators dating the Origination and Collapse of an Asset Price Bubble: Consistency, Convergence Rates and Limit Distributions

Mohitosh Kejriwal, Pierre Perron
Purdue University

This paper studies the asymptotic properties of least squares estimators for the origination and collapse dates of an asset price bubble. The dating estimators are based on an autoregressive model in which the origination of a bubble is associated with a switch from unit root to explosive behaviour while a collapse is indicated by a switch back to unit root behavior. We first show that when the bubble magnitude is fixed, consistency of the estimators only requires the duration of the bubble regime to increase with the sample size. In particular, it is not necessary to assume that the duration is a positive fraction of the sample size, as is typically assumed in the changepoint literature. Further, consistency is derived with respect to the break dates and not just the break fractions. To derive the rate of convergence and the resulting limit distribution, we adopt a mildly explosive representation whereby the bubble magnitude decreases to zero as the sample size increases. Under certain restrictions on the process, we derive a sharp rate of convergence as well as a limit distribution based on which confidence intervals can be constructed. We also show that the dating estimators for origination and collapse are asymptotically dependent so that confidence intervals need to be constructed jointly. This result contrasts with the standard stationary framework where appropriate mixing conditions ensure that changepoint estimators are asymptotically independent. Finally, simulation evidence is presented to assess the accuracy of the asymptotic approximations in finite samples.

10.4 Contributed – General

Thursday 10 September 9.10am – 10.10am

On using predictive-ability tests in the selection of time-series prediction models

*Mauro Costantini, Robert Kunst
University of Vienna*

Comparative ex-ante prediction experiments over expanding subsamples are a useful tool for the task of selecting the best forecasting model class in finite samples of practical relevance. Flanking such a horse race by predictive-accuracy tests, such as the Diebold-Mariano (DM) test, tends to increase support for the simpler structure.

We consider two variants of the DM test, one with naive normal critical values and one with bootstrapped critical values. We compare this selection strategy to a rival one that uses the predictive-ability test by Giacomini-White, which continues to be valid in nested problems and is based on moving windows. Finally, we also consider model selection via the AIC. Our Monte Carlo simulations focus on basic univariate time-series specifications, such as linear (ARMA) and nonlinear (SETAR) generating processes.

10.5 Contributed – Medical – Collating Data

Thursday 10 September 9.10am – 10.10am

Incorporating external evidence on heterogeneity and bias in meta-analysis

Kirsty Rhodes, Rebecca Turner, Jelena Savovic, Julian Higgins
MRC Biostatistics Unit, Cambridge

BACKGROUND: When between-study heterogeneity exists and studies are of high quality, random-effects meta-analysis may be appropriate. However, estimation of between-study variance is difficult in small meta-analyses and study quality sometimes differs substantially. Bayesian meta-analysis allows incorporation of external evidence on heterogeneity and bias. To assist in this, we provide empirical evidence on the likely extent of heterogeneity and effects of bias.

METHODS: Our analyses included some 190 Cochrane reviews that had implemented the Cochrane Risk of Bias tool. Each study had been categorized according to whether reported design characteristics were inadequate, unclear or adequate. To estimate the effects of bias on intervention effects and heterogeneity, we fitted bias models simultaneously to all meta-analyses. Predictive distributions were obtained for heterogeneity expected to remain in future meta-analyses after "removing" bias.

RESULTS: Heterogeneity variances were found to reduce by 24% on average from adjusting for bias due to inadequate/unclear sequence generation, allocation concealment and blinding. In an example meta-analysis, larger studies of higher quality carried greater weight after adjustment for bias based on empirical evidence. Predictive distributions for heterogeneity differed according to the type of outcome under assessment. For a planned meta-analysis assessing a subjective outcome, a predictive distribution for heterogeneity in bias-adjusted intervention effect is log-normal(-3.34, 2.32²). In the example bias-adjusted meta-analysis, incorporating external evidence on heterogeneity led to more precise results.

CONCLUSIONS: Methods to adjust for biases are recommended for use in meta-analyses synthesizing all available evidence. The empirical evidence provided gives useful information on the proportion of heterogeneity due to bias.

10.5 Contributed – Medical – Collating Data

Thursday 10 September 9.10am – 10.10am

A microsimulation model for Health Checks - Evidence synthesis and simulation of what-if scenarios

Arno Steinacher, Anna Goodman, Chris Jackson, Oliver Mytton, Claudia Langenberg, James Woodcock
MRC Biostatistics Unit, Cambridge

The NHS Health checks programme was introduced in 2009 to reduce the burden of cardio-metabolic disease. It aims to identify high-risk individuals by systematically inviting adults in mid-life for a 'health check-up', and then offering appropriate lifestyle advice and/or medical interventions. However doubts about the effectiveness of the programme have been raised because of low uptake amongst those at highest risk and possible poor compliance with treatment. In the absence of direct randomised evidence on the efficacy of this programme, we have developed a microsimulation model. This bottom-up approach allows us to simulate population-level outcomes, for different interventions, based on individual-level data. By combining recent national cross-sectional and longitudinal surveys, routinely-collected data and mortality data, we simulate risk factor trajectories, uptake of Health Checks, subsequent treatments, disease incidence and mortality for individuals over time. An individual-level approach has several benefits: individuals can be assigned multiple continuous variables of interest, which would have required a prohibitive number of states in an aggregate model. The model also allows "what-if" questions, such as changing eligibility criteria or types of intervention, to be readily simulated based on the characteristics of individuals. By doing so, we aim to identify which sections of the populations, such as different socio-economic groups, will benefit most from Health Checks under different scenarios. Uncertainty in the model arising from different sources will be quantified, which may help to prioritise data collection around Health Checks policies in the future.

10.5 Contributed – Medical – Collating Data

Thursday 10 September 9.10am – 10.10am

Application of statistical methods to identify patterns in post-operative complications experienced after cardiac surgery

Rachel L Nash, Barney C Reeves, Gianni D Angelini, Chris A Rogers
University of Bristol

Objectives

In-hospital complications following cardiac surgery are common and can vary in severity. Trials typically report on the occurrence of complications; however multiple complications that patients experience are rarely reported. We aim to identify associations between complications, and their impact on recovery after cardiac surgery.

Methods

Relationships between different complications have been explored in a cohort of cardiac surgery patients; all had participated in a randomised controlled trial. Multiple correspondence analysis (MCA) and latent class analysis (LCA) were used to identify associations and underlying "classes" of individuals based on complication profiles and survival analysis was used to evaluate their impact on recovery time.

Results

Data on 1453 patients from 6 clinical trials were collated. Sixteen complications were investigated; 44% of patients were complication-free after surgery, 31% experienced one complication, 14% two, and 16% three or more. Preliminary investigations showed that patients who experienced more severe complications (e.g. stroke) often had other, less severe, complications as well. Using LCA, three classes were identified; the class labelled 'poor recovery' had a high probability of serious complications (>30%). As expected, post-operative stay was longest in patients assigned to the 'poor' recovery group (median 16 days, 95% CI [13,20] versus 9 [8,10] and 6 [6,6] in the 'moderate' and 'good' recovery groups, respectively).

Conclusions

These analyses have allowed us to identify which complications commonly occur together and those which have the greatest impact on recovery time. Further work will explore how this could be developed to inform an objective measure of recovery after surgery.

10.6 Contributed – Methods & Theory – High Dimensional Data

Thursday 10 September 9.10am – 10.10am

Inference From Noisy, High Dimensional Acoustic Data

*Tim Park, Rakesh Paleja, Matthew Jones, David Randell
Shell*

Distributed Acoustic Sensing (DAS) is a recently introduced sensing technique which has opened up many new possibilities. It is used to measure acoustic and thermal disturbances along the length of an oil producing well. The underlying physical phenomenon is elastic Rayleigh back-scattering of coherent light from inhomogeneities in the core of an optical fibre. The data is collected at a very high temporal frequency and regularly spaced depths. This allows for real time estimation of important quantities such as flow rates, flow composition, slug formation and the propagation of sound waves. These quantities, which previously could only be measured at the surface, can then be used to improve well management and ultimately increase production. Due to the high noise component of the recording we perform our analysis in the Fourier domain. Spectral denoising techniques are then used to improve the signal to noise ratio. Estimating our quantities of interest requires a range of techniques from time series, multivariate and spatial temporal statistics. These include partial least square regression, spline regression, two dimensional Fourier transform and integral transforms. In this talk we present our methodology as well as applications to real wells.

10.6 Contributed – Methods & Theory – High Dimensional Data

Thursday 10 September 9.10am – 10.10am

Estimating Dynamic Graphical Dependencies from Multivariate Time-Series

Alexander Gibberd, James Nelson
University College London

Background

In our data-hungry society we are not only harvesting more data points but also measuring an ever-increasing number of variables. The complex systems represented by such data-sets arise in many socio-scientific domains, such as: cyber-security, neurology, genetics and economics. A dynamic, multivariate approach which models systems jointly can help us focus our analytic and experimental resources on the most significant inter dependencies as they evolve through time. However, such an approach comes with a cost, namely that the number of possible dependency graphs become exponentially large as the number of variables increase. We present and discuss a computationally efficient method that: (i) recovers the key dependency structure within a time-series and (ii) detects structural changepoints where the dependencies between variables appear to change at a systematic level.

Methods

Building on foundations in regularised learning, we formulate a convex relaxed estimator for the partial correlation structure of a piecewise-stationary multivariate Gaussian graphical model. This estimator, which we refer to as the *Group-Fused Graphical Lasso (GFGL)*, can be understood as a semi-Bayesian approach utilising sparsity inducing priors. The output of this model is a dynamic graphical representation of the conditional dependency structure of our time-series data.

Conclusions

We solve the GFGL problem in the MAP setting, develop an efficient proximal splitting algorithm, and demonstrate the estimator performance on synthetic data. Further, we illustrate the potential utility of our method by applying it to pressing real-world problems, such as detecting anomalies in computer networks and examining dependencies within gene activation networks.

10.6 Contributed – Methods & Theory – High Dimensional Data

Thursday 10 September 9.10am – 10.10am

Statistical and Computational Trade-offs in Estimation of Sparse Principal Components

Tengyao Wang, Quentin Berthet, Richard Samworth
University of Cambridge

In recent years, Sparse Principal Component Analysis has emerged as an extremely popular dimension reduction technique for high-dimensional data. The theoretical challenge, in the simplest case, is to estimate the leading eigenvector of a population covariance matrix under the assumption that this eigenvector is sparse. An impressive range of estimators have been proposed; some of these are fast to compute, while others are known to achieve the minimax optimal rate over certain Gaussian or subgaussian classes. In this paper we show that, under a widely-believed assumption from computational complexity theory, there is a fundamental trade-off between statistical and computational performance in this problem. More precisely, working with new, larger classes satisfying a Restricted Covariance Concentration condition, we show that no randomised polynomial time algorithm can achieve the minimax optimal rate. On the other hand, we also study a (polynomial time) variant of the well-known semidefinite relaxation estimator, and show that it attains essentially the optimal rate among all randomised polynomial time algorithms.

11.2 Invited – Official – Recent methodological advances for official statistics

Thursday 10 September 11.40am – 1pm

Model-based inference for official statistics; applications at Statistics Netherlands

Jan van den Brakel

Statistics Netherlands

National statistical institutes traditionally use probability sampling in combination with design-based or model-assisted inference for the production of official statistics. This refers to estimation procedures which are predominantly based on the randomization distribution of the sampling design, while statistical modelling of the observations obtained in the survey only play a minor role. For decades, there has been the prevailing opinion that official statistics must be free from model assumptions, since model misspecification easily translates into wrong statements about the variable of interest. Design-based and model-assisted inference procedures, however, also have some limitations. A major drawback is that they have large design variances in the case of small sample sizes and do not handle measurement errors effectively. In such situations model-based estimation procedures can be used to produce more reliable estimates.

Over the last decade there is an increasing interest at National Statistical institutes to apply model-based inference procedures for the production of official statistics. In this paper several real life examples are presented where model-based procedures are applied to produce official statistics at Statistics Netherlands. One example is the use of a state-space model for the production of monthly figures about the labour force, accounting for discontinuities induced by a redesign of the survey process. A similar approach is currently used to produce quarterly transportation figures. Finally some results of our research into the use of big data as a potential source of official statistics in combination with algorithmic inference procedures known from the area of machine learning will be mentioned.

11.2 Invited – Official – Recent methodological advances for official statistics

Thursday 10 September 11.40am – 1pm

Is it really worth sampling a process more frequently?

Ben Powell, Guy Nason, Paul Smith, Duncan Elliott
University of Bristol

I will discuss recent work in collaboration with the UK Office for National Statistics on assessing the costs and rewards of increasing the sampling rate for a stationary process. The title's question is particularly relevant as opportunities and expectations for fast on-line reporting of official statistics increase. Most of the statistical machinery for constructing a response to the question are already available, but the task remains to connect them coherently. We employ a Bayesian decision-theoretic model as a framework for this statistical machinery, using it to formalize the following idea:

Since our best estimate for the nature of the process in question is based on a finite set of observations, we cannot be totally sure whether increasing the sampling rate or leaving it unchanged ought to lead to lower total costs. Nevertheless, we can still state and justify a preference based on all the knowledge that we do have, including our knowledge of the uncertainty associated with that best estimate.

11.2 Invited – Official – Recent methodological advances for official statistics

Thursday 10 September 11.40am – 1pm

An introduction to applications of wavelet benchmarking with seasonal adjustment

Homesh Sayal, John A.D. Aston, Duncan Elliott, Hernando Ombao
University of Cambridge

Prior to adjustment, accounting conditions between national accounts datasets are frequently violated. Benchmarking is the procedure used by economic agencies to make such datasets consistent. It typically involves adjusting a high frequency process (i.e. quarterly data) so it becomes consistent with a lower frequency version (i.e. annual data). Various methods have been developed to approach this problem of inconsistency between datasets. This paper introduces a new statistical procedure; namely wavelet benchmarking. Wavelet properties allow high and low frequency processes to be jointly analysed. We show that benchmarking can be formulated and approached succinctly in the wavelet domain. Furthermore the time and frequency localisation properties of wavelets allow more complicated benchmarking problems to be considered. Its versatility is demonstrated using simulation studies where we provide evidence showing it substantially outperforms currently used methods. Finally, wavelet benchmarking is applied to official Office of National Statistics (ONS) data.

11.3 Invited – Environment - Spatial preferential sampling and citizen science

Thursday 10 September 11.40am – 1pm

Mitigating the effects of preferentially selected monitoring sites for inference and policy

James V Zidek, Gavin Shaddick, Yi Liu
University of British Columbia

The potential effects of air pollution are a major concern both in terms of the environment and in relation to human health. In order to support environmental policy there is a need for accurate measurements pollution levels. We explore the topic of preferential sampling (PS), specifically where monitoring sites in environmental networks are preferentially located by the designers. We show an example of this using a long standing network in the UK. In the 1960s, over 2000 sites monitored black smoke (BS) air pollution due to concerns about its effect on public health as demonstrated by the famous London fog of 1952. Bayesian models were used to model concentrations over time and space. The large amount of data collected over time means that MCMC may be computationally infeasible and here we use Integrated Nested Laplace Approximations (INLA). Abatement measures led to a decline in the levels of BS and hence a reduction in the number of sites and we show evidence of selection bias with sites being kept in the most polluted areas leading to This causes estimates of metrics used by regulatory agencies to be too high. We present an approach to mitigating the effects of PS. The work has important general implications for the design of monitoring networks and how the information they generate it used. PS has the potential to seriously effect the way in which compliance to regulatory standards is assessed and how the risks that air pollution can have on human health are estimated.

11.3 Invited – Environment - Spatial preferential sampling and citizen science

Thursday 10 September 11.40am – 1pm

The effect of sampling location selection on determining environmental outcomes.

Ron Smith

Centre for Ecology and Hydrology, Edinburgh

Traditionally, sites monitoring air quality were located in areas of expected high pollution. The UK network measuring rural pollutant rainfall ion concentrations was initiated in 1987 as a mixture of historical sites and new locations, giving some loosely defined spatial coverage - a selection process resulting from a mixture of influences. Site numbers have been reduced over the years and a recent study to look at network size highlighted the issue that the mean concentration maps, used for most policy applications, were quite dependent on the number of sites. Further reductions were likely to reduce estimated concentrations in the east of the county and increase them in the west, an example of sampling affecting the spatial pattern as well as the uncertainty.

Ecosystem services provide a value to the benefit humans receive from the environment. Often initial quantification of the services is by relating their delivery to existing maps, e.g. land cover and altitude, so combining remote sensing with localised studies on the quantification of functional relationships. Determining the perceived importance of these estimated services then involves sampling a range of stakeholder opinions: typically locals and visitors, but also there may be remote recipients of the services, and these stakeholders have varying positions of power or influence that may affect their responses. In the age of phone apps, estimates of value are now being obtained from whoever wishes to use the app, or even simply post a photo on the intranet.

The talk introduces some examples of dilemmas related to these sampling issues.

11.3 Invited – Environment - Spatial preferential sampling and citizen science

Thursday 10 September 11.40am – 1pm

Smart study design – optimizing biodiversity monitoring by Citizen Scientists

Stephen Baillie, Mark Miller, Blaise Martay, Robert Fuller
British Trust for Ornithology

Well-designed Citizen Science programmes play a vital role in informing environmental policy, for example through the production of bird population indicators. Smart study design must address both data recording within sample units and the distribution of sampling locations over time and space. For example, the UK breeding bird survey uses a simple line transect protocol that supports models accounting for the effects of local habitat and detectability. Random selection of sample squares within regions allows unbiased sampling while not restricting more detailed coverage where there are many observers. We review examples of such designed monitoring schemes and highlight the strengths and weaknesses of different approaches. Many Citizen Scientists are well able to follow such robust sampling protocols, but excellent volunteer support and feedback are essential.

There is currently much interest in the use of unstructured biological records for ecological monitoring. Such presence-only data raise significant challenges of how best to control for effort and sample coverage. Reliable results can sometimes be obtained through the careful construction of reference samples that control for effort and through the application of techniques such as occupancy modelling. However, comparisons with well-designed sampling schemes suggest that such unstructured data do not always provide reliable results. Furthermore precision per unit effort appears to be much lower than for formal sampling. Semi-structured sampling through portals such as BirdTrack demonstrates that the recording of complete lists is readily promoted amongst birdwatchers and provides a simple yet effective way to control for recording effort within locations.

12.3 Invited – Statistics and the Law

Thursday 10 September 1.50pm – 3.10pm

Statistical evaluation of forensic DNA evidence

Roberto Puch Solis

LGC

Forensic DNA is one of the most powerful evidence types. The technology for producing profiles is continually improving and, at present, DNA systems are very fast and sensitive. A profile can be produced within a few hours and from a fragment of a cell. For example, a profile can be obtained from a fingerprint, i.e. from DNA contained in organic material left when a person touches an object.

The sensitivity of DNA systems posits interpretation challenges for several reasons: many profiles are mixtures, that is profiles coming from more than one person; artefacts of the DNA profiling process (stutters, pull-up); insufficient DNA for producing a person's full DNA profile (allelic dropout); small amount of spurious DNA (dropin) sometimes present in laboratory consumables such as swabs and tubes; and multiple profiles from the same crime sample that need to be interpreted simultaneously (profile replicates).

Part of the evaluation of the DNA evidence in a court of law consists of quantifying the evidential value of a profile obtained from the scene of a crime together with the profile of a defendant. In this talk an inferential framework for incorporating evidential weight in the context of a case being tried at a court of law is discussed. Models that are currently in use in casework and communication challenges are also discussed.

12.3 Invited – Statistics and the Law

Thursday 10 September 1.50pm – 3.10pm

Epidemiological evidence in civil cases.

Jane Hutton

University of Warwick

'If anticoagulants had been administered sooner, my client would not have died.' 'This drug damaged the sight of my patient.' How much money should be awarded to a child who is disabled due to medical negligence? Should a teenager with cancer be given active treatment if doctors estimate he has two weeks to live?

Statements and questions such as these are the basis of civil law suits, in which one party claims damages from a second party, or demands particular actions. Many lawyers still only request expert opinions from medical doctors, although there are many reports on the superior performance of simple extrapolation methods compare to experts, let alone methods which are statistically informed and evaluated.

Statisticians can contribute to civil law suits by finding evidence relevant to the particular case, evaluating it, and then presenting the information, in contrast to a paediatrician who stated in court 'None of my patients is over 18 years, so I do not think this child will live beyond the age of 18.'

Life expectancy for people with neurological injuries is dominated by two groups. In England, expert witnesses cannot be 'bought' by one side. In many other jurisdictions, including Scotland, expert witnesses cannot provide evidence to both sides. If one side instructs all the experts, the other side will be unable to provide an expert's report in response.

In the case of adverse events of drugs, the volume of material can be daunting, and I will explain my strategies for managing this.

12.3 Invited – Statistics and the Law

Thursday 10 September 1.50pm – 3.10pm

The evaluation of evidence consisting of traces of cocaine on banknotes

Amy Wilson, Colin Aitken, Richard Sleeman
Durham University

Mass spectrometry techniques have been developed to obtain measurements of the quantities of cocaine on banknotes that have been seized as evidence for a criminal trial. In this talk a possible framework for the interpretation of these evidential measurements in court is discussed.

There are two particular challenges that will be considered. The first challenge concerns the acquisition of background datasets against which to interpret the measurements of cocaine on seized banknotes. In order to assess a proposition that a seized set of banknotes have been involved with cocaine-related drug crime, it is necessary to obtain a background database of banknotes that have been involved with cocaine-related drug crime. However, this is not practical. Two propositions have been developed to match with existing databases. The advantages and disadvantages of these propositions will be discussed.

The second challenge involves the development of statistical models for measurements of cocaine on banknotes. Cocaine can be found on almost all banknotes in general circulation, making it difficult to distinguish between banknotes that have been associated with crime and banknotes from general circulation. Cocaine is also known to transfer easily between banknotes, so models which account for autocorrelation are required. Three models were developed for the cocaine measurements: a hidden Markov model, an AR(1) model and a non-parametric model. These models were tested and compared to a standard model that assumes independence between banknotes. The results of this analysis will be presented.

12.4 Contributed – Special – P-values debate

Thursday 10 September 1.50pm – 3.10pm

Hypothesis testing, p-values and confidence intervals: Numeracy – good, Comprehension – poor. Comments - definitive guidance required.

*Sam Ellis, Mike Hicks
Defence Ordnance Safety Group, UK*

At the 2014 RSS conference a presentation on the “Lost Lessons of the Masters” described an attempt to avoid the loss of knowledge in the statistical community about the effect of invalid assumptions on parametric testing. Masters such as Pearson, Tukey and Box had quantified these effects but this knowledge has been replaced in teaching and practice by simple rules of thumb and faith in words like “robust”.

A journal has banned hypothesis tests, p-values and confidence intervals in favour of descriptive statistics in the hope of encouraging better research. The RSS website collated responses of which three favoured bans, one was against and one abstained. Someone commented “It is surely the most basic jobs of statisticians to tell us the best way to tell that a result is real or just chance. It’s therefore frustrating to the rest of us to see that this internecine warfare among statisticians still exists.”

As the statistical community grows and many practitioners are not statisticians, they need definitive guidance on the inferences that can be made from the inputs, outputs and the hypotheses being evaluated.

This presentation covers three aspects. One is philosophical and is the source of the “internecine warfare”. The second is technical where a search of the literature indicates that much has been done before but is not understood or used because it is not pulled through into everyday practice. Finally we propose that the RSS takes a view and plays a part in the provision of definitive guidance.

12.4 Contributed – Special – P-values debate

Thursday 10 September 1.50pm – 3.10pm

How good are the reasons given by the editors of Basic and Applied Social Psychology (BASP) to ban hypothesis tests and p-values?

Ian Hunt
EDHEC

This paper argues that BASPs reasoning for banning "statistical inference" is weak.

First, they (the editors) offer a non sequitur: ban p-values because "the state of the art remains uncertain" (2015 editorial) and "there exists no inferential statistical procedure that has elicited widespread agreement" (2014 editorial). I argue inter-subjective (dis)agreement is not decisive.

Secondly, they imply that good inductive inferences require posterior probabilities. This is contentious, especially since both posteriors and p-values are just deductions.

Thirdly, they plead for larger sample sizes "because as the sample size increases, descriptive statistics become increasingly stable and sampling error is less of a problem." This is contradictory: the evidence for this claim is best shown by the statistical inferences being banned.

Fourthly, they correctly assume that with a large enough sample size many "significant effects" (or "null rejections" or low p-values or interesting things or whatever) can be identified by looking at canny descriptive statistics and adroitly drawn charts. But I believe p-values ARE descriptive statistics - with which both frequentists and Bayesians can work.

Finally, BASP "welcomes the submission of null effects". But without tests and concomitant power profiles the evidential value of a "null effect" is unclear.

BASPs editors appear to conclude that modern statistics is inductive and akin to "the art of discovery" (as David Hand puts it). Fair enough. But I conclude that careful deductive inferences, in the form of hypothesis tests with clear premisses and verifiable mathematics, still have a role in discovering interesting things.

POSTER PRESENTATIONS

Please refer to the insert in the conference directory for the final listing with presentation numbers

VECTOR AUTOREGRESSIVE MODEL FOR MONITORING CARBONDIOXIDE (CO₂) EMISSION FROM THE CONSUMPTION OF FOSSIL FUELS

*Olaniyi Mathew Olayiwola, Olawumi Olumayowa Olatunde
Federal University of Agriculture, Abeokuta*

The main source of the Nigerian energy generation is the non-renewable source (which includes the burning of Fossil fuels). Fossil fuels (Coal, Natural gas and Petroleum) are being consumed and certain chemical compounds such as CO₂ are emitted into the atmosphere. The objective of this research is to examine the trend of CO₂ emission from the consumption of Fossil fuels and fitting a model for monitoring the process. Vector Autoregressive Model (VAR) was developed. The Portmanteau test for serial correlation and the Wald test for Granger-causality were carried out. VAR (1) fit the data. The Portmanteau test showed that error term are serially uncorrelated. The Wald test, showed that CO₂ emission from Coal Granger-cause emission from Natural gas and Petroleum and vice-versa. The emission from Natural gas does not Granger-cause emission from Petroleum.

Keywords: Vector Autoregressive Model, Carbondioxide, Granger-causality, Portmanteau test.

The Persistence and Volatility in Naira Exchange Rates returns

OlaOluwa Yaya, Olanrewaju Shittu, Adefemi Adeniran
University of Ibadan

The global financial crisis of 2008 has led to destruction of the structural pattern and volatility persistence of many macroeconomic variables. Particularly, the exchange rates of naira against other currencies around the world have been affected adversely by market reactions. This paper therefore examines these properties in high frequency naira exchange rate time series, before and after the global crisis. Long range dependence techniques and volatility modelling approaches are applied on the series, with absolute and square log-returns as proxies for the volatility series. The significant persistence of volatility in the absolute and square returns of the exchange rates series is observed. The persistence of volatility before the crisis is lower than volatility after the crisis based on the different techniques applied. Further investigation using estimates from volatility modelling confirms the smaller volatility persistence before the global crisis, and possible nonlinear asymmetry is detected during each period. The higher persistence of volatility observed after the global crisis is as a result of the residual impacts of the crisis on the economy that we still experience.

Geeks Corner: how many decimal places to use when reporting risk ratios?

Tim Cole

UCL Institute of Child Health

Objective:

By convention risk ratios are reported rounded to two decimal places; for a ratio of 1.00 this corresponds to a maximum absolute fractional rounding error of 0.5%. However the error is far larger for smaller ratios, e.g. 5% for a ratio of 0.1, and small risk ratios need extra decimal places.

Methods:

A new evidence-based reporting rule is developed, based on a combination of decimal places and significant digits, to ensure that the rounding error averages 0.5% across the range of all possible ratios. The rule is tested on risk ratios reported in BMJ abstracts during 2011-13.

Results:

An optimal cut-off of four was identified, giving a mean rounding error of 0.49% and maximum error 1.3%. The rule of four states: "round the risk ratio to two significant digits if the leading non-zero digit is four or more, otherwise round to three." It uses three decimal places for ratios in the range 0.040-0.399, two decimals for 0.40-3.99 and one decimal for 4.0-39.9 etc.

Among 1250 risk ratios with 95% confidence intervals (CI) published in the BMJ, 89% were reported to two decimal places, 10% to one, and only 0.6% to three. A quarter of ratios and CIs would be reported to other than two decimal places if the rule of four were used, and it would reduce the mean rounding error from 0.76% to 0.49%.

Conclusions:

The rule of four is recommended as the better way to report risk ratios.

Determinants of Health Seeking Behavior in Pakistan: A Complex Health Survey Design

*Shafquat Rozi, Sadia Mahmud, Gillian Lancaster, Wilbur Hadden, Gregory Pappas
Lancaster University*

Background

It is important to understand the health seeking behavior of the population and trend of health services utilization in Pakistan.

Objectives

1. To investigate the determinants of health seeking behavior in Pakistan we suggest a multilevel pseudo maximum likelihood (MPML) approach to estimates model parameters for complex survey design.
2. We will compare the result of MPML with standard multilevel modeling of National Health Survey of Pakistan (NHSP).

Method

A two stage cluster sampling with stratification was employed. Overall 18,315 subjects were interviewed. Health care utilization was considered as a binary outcome. We explored the multilevel modeling approach with sampling weights using two scaling methods.

Results

We found age, gender, marital status, household (HH) ownership of durable goods, urban/rural status, community development index (CDI), and province as significant predictors of health care utilization (p -value <0.05). We also found two significant interactions; between gender and marital status (p -value <0.005), and between the CDI and urban/rural status (p -value <0.045).

The variances of the random intercepts are estimated as 0.135 for PSUs level and 0.224 for HHs. The results are significantly different from zero (p -value <0.05) indicate considerable heterogeneity in health care utilization w.r.t to HHs and PSUs.

Conclusion

Though we observed some divergence in the estimates of slope and variance between un-weighted and scaled weighted analysis but divergences are not marked. This may have occurred because of larger cluster size and relatively small ICC. Results suggest that more health centers should be set up in rural areas and introduce gender sensitive programs.

Peer pressure and family smoking habits influence smoking uptake in male adolescents attending public and private schools in Karachi, Pakistan: multilevel modeling of survey data

Shafquat Rozi, Sadia Mahmud, Gillian Lancaster
Lancaster University

Introduction

Adolescents spend a considerable amount of their time in school, the school environment is therefore important for child outcomes. Among young teens, about one in five smokes worldwide.

Objectives

- To develop two level random effects logistic regression model and GEE model to identify predictors of smoking on teenage children attending school and the results will be compared with a conventional logistic regression model.
- To develop random coefficient model to assess if the variability between schools is different for the public and private.

Methods

A two-stage cluster sampling with stratification was employed. We interviewed 772 male secondary school students. The outcome variable is smoking status of the students. We have two level data with single level of clustering.

Results

The results showed that age, mother's education and parents, family and friends smoking all contributed to teenagers smoking (p -value <0.05). Between cluster variance is significantly different from zero (p -value of likelihood ratio test = 0.01), which indicates that there is variability between schools. The ICC is greater than zero (ICC =0.15).

There is variability among schools but it is not different for public & private (p -value > 0.99).

Conclusion

Random effect model and GEE take correlation in to account in the inferential process and indicating that there is variability between schools and we need to take cluster variation in to account using multilevel modeling. The results point out the need for an effective tobacco control program among adolescents. Parental counseling about influence of family tobacco use on children may bring about encouraging results.

Bayesian Criteria for Model Choice for Hidden Markov Models

Safaa Kadhem

University of Plymouth

We discuss the problem of Bayesian model choice in the context of hidden Markov models. We first introduce criteria proposed from the Bayesian point of view, which include the modified AIC, the modified BIC, the DIC and the WAIC. These criteria have been proposed for model choice for several models where the independence assumption between observations holds. In HMMs, the independence assumption between observations is violated. As a result, the application of these criteria for models with sequential data, i.e. HMMs, is problematic. Breaking the dependence assumption between observations in HMMs, by assuming that the observed data are conditionally independent given the hidden process, could possibly allow the application of those criteria for HMMs. The DIC and WAIC are essentially based on a Bayesian deviance. We propose modified versions for AIC and BIC, that are based on a Bayesian deviance instead of the frequentist deviance. The approximate deviance in the context of HMMs is based on the complete data log-likelihood function which is recursively evaluated by integrating out the hidden states using the forward-backward recursion. Simulation studies on both synthetic and real data sets have been used to compare the performance of these criteria for several competing HMMs.

On New Methodology for Sufficient Dimension Reduction

Andreas Artemiou
Cardiff University

In this talk we will present new results on new methodology for sufficient dimension reduction (SDR). SDR methodology performs supervised dimension reduction in regression and classification problems. The main effort is focused on performing the dimension reduction without losing information for the conditional distribution $Y|X$.

The methodology we present here is based on ideas that were already introduced in the SDR framework. It uses the ideas of slicing, inverse moments and SVM to perform SDR. The new methodology provides more robust estimation to the presence of inliers and outliers and gives new insights on the use of machine learning techniques in the SDR framework. It can be also shown that the new method can be seen as the result of a weighted average between two existing methods, thus providing the framework for Flexible Dimension Reduction based on a tuning parameter.

A meta-analytic generalisation of the Lincoln-Petersen-Chapman estimator for mark-resight experiments

Dankmar Böhning

University of Southampton

The Lincoln-Petersen and Chapman estimators are the most frequently used approaches for estimating a population size. The work is motivated by mark-resight studies in stray dogs and cats in south-eastern Europe. Typically, the animals are marked (by some colour spray) without being actually caught at some first occasion. At several later occasions it is noted whether the observed animal carries a mark or not (again without catching them). This leads to a series of Lincoln-Petersen or Chapman estimates of the population size. The difference between mark-resight experiments and mark-recapture experiments is the fact that in mark-resight experiments it is not recorded at which of the resight occasions an animal is observed. If there is only one resight occasion mark-recapture and mark-resight experiments are identical. Typically, mark-resight experiments are analysed by presenting a series of Lincoln-Petersen and Chapman estimates. We propose to view this as a meta-analytic setting and suggest summary estimators that provide a single estimate of the population size as a more efficient approach. In particular, we suggest a Mantel-Haenszel type estimator which is more robust in sparsity situations (the frequency of those seen at first and later occasions can become small). Simulation work shows that the Mantel-Haenszel type generalisation of the Lincoln-Petersen and Chapman estimator behaves well in terms of bias and variance. In addition, we consider variance estimation which is more complex in this situation as identical animals might contribute several times to the summary. A parametric bootstrap procedure is proposed for variance estimation in this case.

House Prices: Innovating Small Area Official Statistics using Open Data

Nigel Henretty, Frank Donnarumma
Office for National Statistics

In February 2015, the Office for National Statistics published House Price Statistics for Small Areas. Using Open Data from the Land Registry, this provides the clearest possible picture of the average value of house sales on an annual basis going back to 1995.

We took responsibility for the production of these statistics from the Department for Communities and Local Government, and brought together data for a range of different geographies in one useful suite of house price data, designed to be equally accessible to both the expert user and the inquiring citizen. Now, with 19 years' worth of data including 3 geographies, 5 measures and 5 house types, there are over 62 million data points on housing to explore!

In our presentation, we will discuss the innovative approach we took to Open Data, user engagement and branding to create official statistics with impact and longevity.

Simultaneous Confidence Sets For Several Effective Doses Via The Inversion Of Confidence Bands Method

Daniel Tompsett, Stephanie Biedermann, Wei Liu
University of Southampton

Confidence sets for effective doses can be constructed using the inverse dose-response curve of 1-alpha simultaneous confidence bands on the maximum likelihood estimated, asymptotic logistic regression model, known as the inversion of confidence bands method. Confidence sets of this form are notably overprotective of the nominal 1-alpha percentage coverage when simultaneous confidence sets are sought for two, three, or several effective doses at once.

In this research, simultaneous confidence sets are constructed specifically for two, three, and any finite number k effective doses, via the creation of specialised simultaneous confidence bands. Simultaneous confidence sets have been fully specified for two effective doses for a multivariate logistic regression model using a method involving Sidaks inequality. Furthermore, fully specified simultaneous confidence sets are demonstrated for three effective doses for a univariate logistic regression model, with a method, using theory involving vector projection outlined to establish similar sets for any number k effective doses. A working example is given to illustrate that specialised confidence sets of this type are less conservative than the currently established inversion of confidence bands method, providing simultaneous coverage closer to the nominal 1-alpha level, though the improvement deteriorates as more sets are sought at once.

Quantification of the risk of cervical cancer in symptomatic primary care patients

Sarah Walker, Chris Hyde, Willie Hamilton
University of Exeter

Background

In the UK approximately 3,100 women are diagnosed with cervical cancer each year. It is the most common cancer in women under 35. Over a quarter of cases are diagnosed symptomatically, rather than through screening; especially in young women. Suspicion of cancer and early recognition of symptoms by both patients and GPs are essential if delays in diagnosis are to be reduced, especially in women who have not been screened. This study will identify and quantify the features of cervical cancer in symptomatic primary care patients to aid selection of patients warranting further investigation.

Method

Women aged ≥ 40 years diagnosed with cervical cancer between 2000-2009 and up to five age, sex and practice-matched controls were selected from the Clinical Practice Research Database. Features of cervical cancer will be identified in the year pre-diagnosis, and odds ratios calculated using conditional logistic regression. Positive predictive values (PPVs) will be calculated for consulting women.

Results

1,006 cases and 4,992 age, sex and practice-matched controls are available for study. Symptom, sign and laboratory test analyses are underway; available prior to Sept 2015. We expect most symptoms reported from secondary care studies will also be strongly associated with cervical cancer in primary care.

PPVs will quantify the risk of cervical cancer for individual (or pairs of) symptoms commonly associated with cervical cancer in a woman consulting with her GP. Where the symptom profile may warrant investigation but risk of cancer is low, the risk is quantifiable and may offer some comfort to the patient.

Comparison of statistical algorithms for the detection of infectious disease outbreaks in large multiple surveillance systems

*Doyo Enki, Paul Garthwaite, Paddy Farrington, Angela Noufaily, Nick Andrews, Andre Charlett
Plymouth University*

A large-scale multiple surveillance system for infectious disease outbreaks has been in operation in England and Wales since 1991. Changes to the statistical algorithm at the heart of the system have been proposed and the purpose of this work is to compare two new algorithms with the original algorithm. Test data to evaluate performance are created from weekly counts of the number of cases of each of more than 2000 diseases over a twenty-year period. The time series of each disease is separated into one series giving the baseline (background) disease incidence and a second series giving disease outbreaks. One series is shifted forward by twelve months and the two are then recombined, giving a realistic series in which it is known where outbreaks have been added.

The metrics used to evaluate performance include sensitivity, specificity, time before detection and the relative size before detection. Also, a scoring rule is used as a measure of performance that appropriately balances sensitivity and specificity. The scoring rule was chosen to be sensitive to variation in probabilities near 1. In the context of disease surveillance, a scoring rule can be adapted to reflect the size of outbreaks and this was done. Results indicate that the two new algorithms are comparable to each other and better than the algorithm they are designed to replace.

The Network Structure of R Packages

Andrie de Vries, Joseph Rickert
Revolution Analytics

Over the past decade, the R programming language has moved from being a rather specialized platform supporting statisticians to a widely used tool at the center of the new developments in data science.

Much of the growth in the capabilities of the R language is due to the success of R's "package" system. R developers add new capabilities to the language by writing functions that are grouped together into packages and submitted to repositories where they are tested and made available to R community. CRAN (> 6,000 packages) and BioConductor (> 900 packages) are the two primary repositories. These repositories are managed independently with different release cycles and different conformance policies.

In this paper, we use statistical network theory and the algorithms implemented in various R packages including igraph and miniCRAN to analyze and visualize the connectivity structure of packages in CRAN and BioConductor. We also test the hypothesis that the different management policies of the two repositories are reflected in the properties of their graph networks.

Modelling MRSA incidences in surgical wards

Xing Lee, Tony Pettitt
Queensland University

Healthcare associated infections, particularly those involving multidrug-resistant organisms (MRO), are associated with increased mortality, morbidity and cost to both patients and hospitals worldwide. Preventing such infections is imperative, particularly in light of the substantially slower development of new antimicrobials to treat these infections.

One avenue of investigation is to use statistical models to further insight into MRO transmission mechanisms within a hospital ward, supplementing more conventional statistical analyses and adding value to existing data. This concept is used to further investigate the impact of an extra cleaner on MRSA incidence rates of surgical wards in a UK hospital (Dancer et al., (2009). Measuring the effect of enhanced cleaning in a UK hospital: a prospective cross-over study. *BMC medicine*, 7(1), 28).

A non-homogeneous Poisson process model was developed to quantify the relative contributions of colonised patients and environmental contamination toward the observed MRSA incidences. Differences in these relative contributions during the periods with and without the extra cleaner provide a quantitative measure of how this intervention impacted the two different transmission routes. Imputation and smoothing methods were used with the imperfectly observed data to estimate model parameters.

Despite the small number of cases observed during the study period, the model parameters were able to be estimated and differences between the periods were highlighted. In particular, we showed how the simple cleaning intervention affected the relative contributions of the different transmission routes.

Logistic Regression and Predictive Analytics: A role in identifying probabilities of student success

Carol Calvert
Open University

The objective of this work was to determine whether the vast store of data available to the University could be used to help tailor support to students. Dependent on that evidence being identified the subsequent aim was to develop and implement a sustainable approach to integrating the information within the University support systems to its students.

A logistic model, developed using SAS E Miner software, was built to generate the probabilities that students passed certain milestone in their academic journey. A core set of variables, were identified which proved robust over time. The variables were selected from hundreds of data items either routinely collected, or derivable from them, on hundreds of thousands of students over several years. The data mining techniques of SAS E miner were exploited although eventually a basic logistic regression model proved as effective as more complex options.

The identified variables were then used to predict the number of students who were likely to leave at different milestones. The predictions were sufficient robust to base interventions for certain students at specific time points.

Model Based Approach for Determination of Robust Optimum Plot size and shape under Field Condition

Satyabrata Pal

Honorary Visiting Professor, Indian Statistical Institute

Determination of optimum plot size has long been a very important area of study to the agriculturists and statisticians since the first contribution on it appeared in the paper by Smith (1938), and worth mentioning is that precise solution to it is yet to come. Interestingly, further researches/methods (in Faground and Meirvenne, 2002 and others) do not take care of the type of correlation-structure existing in real-life field data. This paper considers the model as: the uniformity trial data $Y(\mathbf{s})$ on a spatial location \mathbf{s} are written as: $Y(\mathbf{s}) = \mu + e(\mathbf{s})$, $V(Y(\mathbf{s})) = V(e(\mathbf{s})) = \sigma^2$; $\text{Cov}(Y(\mathbf{s}), Y(\mathbf{s} + \mathbf{h})) = \rho^{||\mathbf{h}||} \sigma^2$, here values of ρ equal to 0.1, 0.2, ..., 0.5, in case of data from the designed experiments, variogram of the residuals are to be modelled in the above manner. The expressions of the theoretical variograms, $2\gamma(\mathbf{h})$ (under the above model) have been obtained for Plot sizes, $l \times k$ ($l = 1, 2, 3, \dots, k = 1, 2, 3, \dots$), etc. (area of plots 25 squared units). The variogram-graph-plots are constructed for each of the admissible plot sizes. For each plot size the point h_{opt} is determined (for which point, (i.e., at h_{opt}), the value of the radius of curvature, r_c , is minimum). Any one of the plot sizes, 2x5, 2x6, 3x5 and 3x6 is the robust optimum plot size irrespective of the values of ρ .

Modelling Total Electricity Generation in Nigeria using Response Surface Methodology

Oluwaseun Otekunrin, Oladele Ariyo
University of Ibadan

The objective of this study was to model and validate total electricity generation in Nigeria using Response Surface Methodology (RSM). The factors considered were Population/Million (POP), Tariff/Naira (T), Dry Natural Gas Consumption/Billion kWh (NGC) and Hydroelectric Power Consumption/Billion kWh (HPC). Total Electricity Generation/Billion kWh (TG) was the response variable. Secondary data which covered a period of 1985 to 2014 was used in the study.

The lack-of-fit of the first order model was significant at 5% level ($p = 0.000002695$) necessitating the need to move to the second order model. The non-significant lack-of-fit of 0.1001209 at 5% level in the second order model implied that the model was all right for prediction with multiple R^2 of 0.9842.

Using the stationary point technique, the predicted response (TG= 33.5 Billion kWh) was optimized at levels 31.79 naira, 556.3348 Billion kWh, 21.2697 Billion kWh and 174.8507 million people for T, NGC, HPC and POP respectively. This optimum response (TG=33.5 Billion kWh) exceeded the current maximum TG in Nigeria (28.4 Billion kWh in year 2013). For model validation, actual values of TG in Billion kWh (26.5, 28.4, and 27.78) were relatively close to their predicted values (25.7, 27.97, and 27.42) for years 2012, 2013 and 2014 respectively.

Total electricity generation in Nigeria was modeled using RSM. The developed model has good predictive ability. The optimum value obtained for TG showed that Nigeria has not reached the required level for total electricity generation.

Statistical calibration for infinite many future values and exact simultaneous tolerance intervals for linear regression

Yang Han
University of Exeter

Objectives

Simultaneous tolerance intervals (STI's) can be used to construct prediction intervals for the future values of the response variable corresponding to an unlimited sequence of observations of the covariate variable. They can also be used for multiple-use calibration which employs unlimited observations of the response variable to construct confidence intervals for the corresponding covariate variable, based on the same training data set. Wide range of applications of multiple-use calibration motivates people to construct STI's. The only exact STI's published in the statistical literature are provided by Mee, et al. (1991) and Odeh and Mee (1990). But they are for a multiple linear regression model, in which the covariates are assumed to have no functional relationships. When applied to polynomial regression, the resultant STI's are conservative. The main purpose of this research is to construct exact STI's over any given covariate interval, and to compare them under the average width criterion.

Methods

Methods given in this paper allow the construction of exact simultaneous tolerance intervals over any given interval. We illustrate the proposed methods with a real example.

Results and conclusions

Calibration for infinitely many future y -values requires the construction of STI's. We are able to construct exact STI's for both polynomial regression and multiple regression. The exact STI's presented in this paper are narrower than STI's available in the literature.

Comparison of Models and Priors in Bayesian Subgroup Analysis Using WAIC

Jun Takeda

Astellas Pharma Inc.

In a Bayesian subgroup analysis, selection of the model and the prior is quite important because they affect the extent of shrinkage appeared in the posterior distributions of the subgroups. We suggest a framework to explore (semi) optimal models to a data set based on a model fitting criterion.

Firstly we define a set of model families. Such a set includes the common effect model (all subgroups have the same mean), the hierarchical models, and the individual model (all subgroups have independently different means). We suggest that the hierarchical model family includes different kinds of prior distributions with various ranges of hyper parameters for weak, moderate, and strong shrinkage. In addition we suggest that such a set of families also include models that continuously connect between the common effect model and the individual model. Secondly WAIC (Widely Applicable Information Criterion or Watanabe-Akaike Information Criterion) is calculated for the posterior distribution of each model in the model family set. Finally the values of WAIC are compared to consider which models/priors well fit the data.

We applied the above procedure to data sets for multi-regional clinical trials and for some other settings. The models with smaller WAIC, which are considered as better models, range from those with strong shrinkage to almost no shrinkage depending on the data set. A finding is that some data sets show relatively smaller values of WAIC for models lay between the common effect model and the individual model.

Receiver Operating Characteristic Analysis for Evaluating Diagnostic Performance of Urine albumin concentration and calculation of albumin-to-creatinine ratio to screen hypertensive patients for persistent albuminuria: Findings from community based study of Karachi Pakistan.

Muslima Ejaz
Aga Khan University

Albuminuria is one of the earliest markers of kidney damage and screening for albuminuria is recommended in patients at increased risk for chronic kidney disease. Accurate estimation of albuminuria requires timed collection of urine over 24 hour, which is cumbersome in routine clinical practice. An alternative method is detection of albuminuria in spot urine sample, either by urine albumin concentration (UAC) alone or albumin-to-creatinine ratio (ACR), which are relatively easy to obtain in the ambulatory setting.

Methods:

A community based cohort of hypertensive subjects (n=1340) were evaluated for the presence of persistent albuminuria, defined as UAC of 20mg/L and ACR of ≥ 30 mg/g of creatinine persisting for more than 3 months. Diagnostic performances of UAC and ACR for albuminuria are expressed in terms of specificity, sensitivity and areas under receiver operating characteristic (ROC) curves.

Results:

The overall prevalence (95% CI) of persistent abnormal albuminuria, calculated by UAC and ACR was 9.3% (7.8% - 10.8%) and 8.1% (6.6% - 9.4%) respectively: 12.8% in men and 7.2% in women (P=0.04), 10.8% in men and 7.4% in women (P=0.04) respectively. The areas under ROC curves for UAC were 0.86 (0.82-0.90) and 0.88 (0.84-0.92), and areas under the ROC curves for ACR were 0.86 (0.82-0.89) and 0.90 (0.86-0.93), respectively, in women and men. For UAC, the sensitivity and specificity were 37 and 97%, respectively, in women and 69 and 94%, respectively, in men at the conventionally recommended value of 20 mg/L.

Conclusion:

Both UAC and ACR are acceptable tests for population screening for albuminuria.

Using difference in differences, interrupted time series to evaluate the impact of the prescription fee abolition in Scotland

*Andrew James Williams, Ruth Jepson, Marion Bennie, Chris Dibben, Anne Ludbrook, Jill Pell, John Frank
University of Edinburgh*

In April 2011 the Scottish Government followed Wales and Northern Ireland in abolishing fees for prescriptions following a 3 year period of stepped fee reduction. In this study of those aged 19-59 years, the objectives were to identify whether these changes resulted in: a) increases in the number of prescriptions for inhaled corticosteroids; b) reductions in hospital admissions for asthma or chronic obstructive pulmonary disease (COPD) when compared to those with a condition (diabetes) or in an age group for whom prescriptions had historically been free. These groups were identified as being most likely to demonstrate any effect. The Scottish Prescribing Information System (PIS) holds data on community issued and dispensed prescriptions. Limitations of the PIS meant that only practice-level data and secondary non-adherence could be assessed; however, linking to the Scottish Morbidity Record (SMR01) permitted a natural experiment.

A difference in differences interrupted time series approach was taken to analyse data from 798 practices across Scotland for the period from July 2005 to December 2013. Across the intervention and control groups the number and cost of prescriptions markedly increased over the period of reduction and abolition. The trend in admissions for diabetes did not change, whereas admissions for asthma/COPD levelled-off in both the intervention and age-based control groups following the reductions. The analyses were stratified by deprivation and other practice characteristics; with results indicating that the policy change may have affected hospital admissions. The nature of the previous fee exemption system makes deducing the potentially attributable effect difficult.

A state space model for the UK Labour Force Survey

Duncan Elliott, Ping Zong

Official for National Statistics

The UK Labour Force Survey publishes estimates of employment and unemployment for the UK each month. These estimates are often referred to as rolling quarterly estimates as they relate to the most recent three month period. Previous interest in a monthly estimate led as a compromise to the publication of the rolling quarterly estimates, but there continues to be interest in the production of monthly estimates for which there is currently no National Statistic. We present a state space model that deals with aspects of the survey design, in particular survey error autocorrelation and potential rotation group bias that could provide monthly estimates of variables in the LFS with no changes required to the survey design. Monthly wave specific estimates are used as observations in a multivariate state space model to derive improved monthly estimates over the current experimental monthly statistics. We evaluate the potential to use these model based estimates to replace the current experimental monthly statistics.

Bias assessment and reduction for the 2SLS estimator in General dynamic simultaneous equation models

Dandan Wang
Cardiff University

We consider the bias of the 2SLS estimator in general dynamic simultaneous equation models (DSEM) with g endogenous regressors. By using asymptotic expansion techniques we approximate 2SLS coefficient estimation bias under innovation errors, p lagged-dependent variables and strongly-exogenous explanatory variables. Large- T approximations bias of the structural form is then used to construct corrected estimators for the parameters of interest in the general DSEM (C2SLS). Simulations show that the C2SLS gives almost unbiased estimators. Alternatively, the numerical bootstrap method results suggest that the non-parametric bootstrap could be used in 2SLS for improving estimation in general DSEM.

Childbearing intentions of people living with HIV/AIDS and are on Antiretroviral Treatment in Uganda

Dorothy Mubuuke

Makerere University, Uganda

At the start of the HIV/AIDS pandemic, reproduction by PHAs was roundly discouraged because of the risk of sexual and vertical transmission of HIV. ART has now significantly reduced both these risks. This paper examined childbearing intentions by PHAs on ART and identified factors associated with it. The paper used quantitative and qualitative data collected from randomly selected PHAs using a structured questionnaire and focus group discussions respectively. Chi-square and stepwise binary logistic regression were used to analyze quantitative data while the master sheet method was used to analyze qualitative data. The results showed that 30% of the PHAs of which 33% were males and 28% were females, reported intentions to bear children. Females: persons older than 29 years; persons with one or more children; the widowed; separated/divorced persons; Catholics and women engaged in household chores were less likely to report intentions to bear children. Conversely, younger, childless never married and PHAs engaged in business reported intentions to bear children. This finding shows that ART has the potential to make PHAs review their reproductive intentions. It calls for measures to protect the reproductive rights of PHAs while at the same time, prevent the spread of HIV.

Key words: childbearing intentions, ART, PHAs, Uganda

Reducing road worker risk

Caroline Wallbank

Transport Research Laboratory

Road workers are essential to the smooth running of the road network; however the job is not without its risks. One of the highest risk activities for road workers is exposure to live traffic whilst deploying and removing signs, particularly when placing signs in the central reservation as road workers have to carry equipment across the carriageway.

In 2009, the Highways Agency set out its Aiming for Zero Road Worker Safety Strategy with the aim of achieving a substantial reduction in the need for road workers to cross live carriageways on foot. Eliminating the need for signs in the central reservation is consistent with this aim. An on-road trial to evaluate driver behavior on approach to road works with nearside-only signing was commissioned.

A trial was designed to compare driver behaviour on approach to road works signed on both sides of the carriageway (the control layout) to driver behaviour on approach to road works with signs on the nearside of the carriageway only (the experimental layout).

The trial was a success and the removal of offside signs was approved as a technique for single lane closures. Since this first trial, the Highways Agency and its service providers have funded research to trial the technique for other closure types (e.g. Lanes 1 and 2). Use of this signing layout has resulted in a substantial reduction in carriageway crossings and subsequently in the risk experienced by road workers on the network, with no increase in risk to the road users.

Predicting the impact of vehicle safety developments in emerging markets

Caroline Wallbank, Louise Lloyd
Transport Research Laboratory

Vehicles sold in emerging markets are not typically regulated to the same extent as seen in industrialised regions. Casualty rates are considerably higher in these emerging markets, and the lack of vehicle safety regulation is responsible for at least some of the difference. With rapid growth in passenger cars expected, the number of road deaths and casualties in emerging markets is likely to rise, unless targeted interventions are planned and initiated.

The objective of this work was to quantify the casualty benefit that could be realised in Malaysia, if the experiences and lessons learned in Europe, including minimum car safety standards and consumer testing, were applied. Clear differences in vehicle safety developments in Europe compared with Malaysia were identified, and suggest that new cars sold are approximately 10 years behind today's equivalent European cars in terms of vehicle safety developments.

This work involved logistic regression modelling to calculate the casualty benefit experienced in Europe due to improvements in vehicle safety and application of these savings to the emerging market, taking into account the likely trend in casualties and exposure as the use of passenger cars increases. As a result, the estimated potential saving is between 1,200 and 4,300 Malaysian fatalities by 2030.

Relationships between attributes of forecast performance

Keith Mitchell, Christopher Ferro
University of Exeter

For an event with only two possible outcomes, a probability forecast is a statement about how likely it is that the event will occur. How good a particular probability forecast is, can be quantified by using a scoring rule, which takes a probability forecast and corresponding event outcome and returns a value that conveys the closeness of the probability forecast to the outcome. The performance of a forecaster is given by the expected value of the scoring rule over the joint distribution of the probability forecast and the event. To assess a forecaster, the expected value of the scoring rule may be decomposed into components, each of which represents a particular quality of the forecaster. There are two main decompositions: the uncertainty-resolution-reliability (URR) decomposition and the refinement-discrimination-discernment (RDD) decomposition. For a binary event, the decompositions are particular cases of more general results -- Brocker (2009)* and a comparable result we present here. Moreover, for the popular Brier scoring rule, we show that the two decompositions are intimately related; the link between the two decompositions is another tool for analysing probability forecasts of binary events - the reliability curve. This reinforces support for the reliability curve as a fundamental tool for analysing probability forecasts of binary events.

*Brocker, J. (2009), 'Reliability, sufficiency, and the decomposition of proper scores', Quarterly Journal of the Royal Meteorological Society, Volume 135, 1512-1519.

Discrete Weibull Regression Model for Count Data

Hadeel Kalktawi, Keming Yu
Brunel University London

Data can be collected in the form of count data in many situations. Poisson regression can be considered as the most common model to analyse the effect of covariates on count but subject to the constrain of equi-dispersion. Then, the negative binomial regression is a good alternative for count data analysis with over-dispersion, but it cannot be the best choice to model the under-dispersed data. In this study, the discrete Weibull regression model allowing for over- and under-dispersion, is introduced.

A statistical approach on determination of ETL for pests on crop

Arunava Ghosh, Satyabrata Pal
Uttar banga Krishi Viswavidyalaya, India

Pest attack creates enormous damage to crops and poses challenging task to agricultural scientists all over the world. Economic threshold level (ETL) is an important component in pest management and control. Usually, it is determined by the grower/technologist utilizing his experience on a crop. Knowledge of ETL helps reduce crop loss (and ensure less pesticide application), and as a consequence, profit is increased. Also substantial knowledge is required on the dynamics of the pest population, in order to determine the density at which the economic injury level (EIL) may be prevented (Weersink et al. 1991). This paper is devoted to the development of an analytical method (probabilistic) for determination of ETL, which is defined as the density at which control measures should be determined to prevent an increasing pest population from reaching the economic injury level. Occurrence probabilities of the number of pests have been computed from best fitted probability distribution on pest infestation data. Semi parametric and non parametric models have also been applied to find out change of pest infestation pattern over time. The above method is demonstrated on real life data set on major pests (whitefly and blackfly) incidence on betelvine, obtained from an experiment designed for that purpose. The findings emerged from the study reveals that ETL values of two pests white fly and black fly lies in the range, (3-4) and corresponding cumulative probability range is 0.25 to 0.45.

Key words: Economic Threshold Level (ETL), Economic Injury Level (EIL), Nonparametric and semi-parametric models, Occurrence Probability.

Assessment of the Impact of climate change on the Health Sector in Uganda: A case of Malaria and Cholera epidemics

Dorothy Ndagire

Makerere University, Uganda

Health which is directly and indirectly impacted upon by climate change is defined as a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity. The paper intends to assess the impact of climate change on human health in Uganda focusing on malaria and cholera. Available data on malaria and cholera from Health Management Information Systems and Epidemiological Surveillance Division was used to get literature review, a questionnaire was conducted in 12 districts that were purposively selected because of their past vulnerability to climate change. Data was entered and analyzed using MS Excel. From the finding it's clear that malaria and cholera outbreaks and epidemics are partly linked to climate change. Climate change however, is not the only factors explaining the occurrence of malaria and cholera epidemics. Other factors like increasing resistance to anti-malarial drugs, weak health system, late treatment, costly preventive interventions, high prevalence of low quality treatment outlets both formal and informal combined with poor sanitation, unhygienic food contribute significantly. These factors have led to loss of lives, economic loss and destruction to infrastructure. Therefore Health sector must collaborate with other sectors in order to have an effective adaptation mechanism.

Modeling heteroscedasticity of Albanian exchange rate time series in R software and efficient estimates of Beta.

*Kristel Bozhiqi, Mirjeta Dëra (PASHA), Ranela Kapçiu
Aleksandër Moisiu University, Albania*

The ability to capture non-linear dynamics in real life situations especially in financial markets has made GARCH models very useful. This paper discovers those heteroscedasticity behaviours in exchange rate time series and tries to approach the best model. After we have fitted the best ARMA model based on information criteria (AIC, BIC and SIC) we show in the residuals of ARMA models the heteroscedasticity of this time series. Heteroscedasticity presence prompts us to fit data with GARCH(1,1) and extensions such as E-GARCH(1,1) and I-GARCH(1,1) which modeled in R generate themselves the Beta coefficient. To find the best Beta we test different Beta values around those generated beta. We do these simulations and compare the results with original exchange rate data.

Models used: GARCH(1,1), E-GARCH(1,1), I-GARCH(1,1)

Software used: R

Keywords: heteroscedasticity, GARCH, exchange rate, model, Beta.

A comparison of model selection and model averaging for prediction using multiply-imputed datasets

Khuneswari Gopal Pillay, John McColl
University of Glasgow

Model selection and model averaging become more complicated in the presence of missing data. The aim of this paper is to compare model selection and model averaging on multiply-imputed datasets, when the primary purpose of the analysis is to predict the value of a response. The performance of three strategies (inclusive, restrictive and non-overlapping) for building imputation and prediction models is discussed in the context of an analysis of datasets from the Gateshead Millennium Study. Imputation was carried out using chained equations (via the "norm.nob" method in the R package MICE). Possible combinations of linear multiple regression models for prediction were fitted and the best model was chosen using the model selection criterion AICc (corrected AIC). A modified version of the STACK method (M-STACK) was used for model selection with multiply-imputed datasets. Non-Bayesian model averaging was explored to average the estimates of multiply-imputed datasets using AICc based weights. The performance of M-STACK and model averaging was compared using mean square error of prediction (MSE (P)) in a 10% cross-validation test. The results showed that the M-STACK provided better prediction than model averaging. The inclusive strategy for building imputation and prediction models was better than the restrictive and non-overlapping strategies. The presence of highly correlated covariates and response is believed to have led to better prediction. Therefore, it is advisable to use M-STACK with an inclusive model-building strategy and highly correlated covariates to make predictions in the presence of missing data.

Overview of the national accounts

Pete Lee

Office for National Statistics

Would you like to know more about the National Accounts? To understand what they are, how are they put together and why they matter to everyone in the UK?

This presentation will cover some of the key economic statistics, e.g. gross domestic product, gross national income, balance of trade, real household disposable income, etc.

Combining information from multiple flood projections in a hierarchical Bayesian framework

Nataliya Le Vine

Imperial College London

The study demonstrates the potential, in the context of flood frequency analysis, of a recently proposed hierarchical Bayesian approach to combine information from multiple models. The approach facilitates explicit treatment of shared multi-model discrepancy as well as the probabilistic nature of the flood estimates, by treating the available models as a sample from a hypothetical complete (but unobserved) set of models. The advantages of the approach are: 1) to insure an adequate 'baseline' conditions with which to compare future changes; 2) to reduce flood estimate uncertainty; 3) to maximise use of statistical information in circumstances where multiple weak predictions individually lack power, but collectively provide meaningful information; and 4) to adjust multi-model consistency criteria when model biases are large. The findings are of value with regards to how conclusions about changing exposure to flooding are drawn.

Graphical least squares estimation

Saeed Aldahmani, Hongsheng Dai
University of Essex

Objectives:

A new method, named as Graphical Least Squares, is proposed for solving the linear regression problems when the number of observations n is smaller than the number of predictors v .

Models:

The proposed method is based on the idea of graphical models and provides unbiased parameter estimates under certain conditions, whereas the existing methods such as ridge regression, LASSO and least angle regression (LARS) give biased estimates. The new method is aimed to provide a detailed graphical correlation structure for the predictors, so that the real causal relationship between predictors and response could be identified. In contrast, existing methods often cannot identify the real important predictors which have possible causal effects on the response variable.

Conclusion

To evaluate the proposed method, real and simulated data sets are used and the results are compared with ridge regression, LASSO and LARS. It is revealed by our experiments that the proposed method is better than all the other methods, when $n < v$.

The usefulness of statistics for understanding uncertainty in environmental models

*Lindsay Lee, Carly Reddington, Kirsty Pringle, Graham Mann, Alex Rap, Ken Carslaw
University of Leeds*

All models of environmental systems are uncertain because the model design cannot be a faithful reproduction of reality and the magnitudes of many model quantities (parameters) are poorly defined by theory or experiment. Understanding how uncertainty in the model parameters affects their outputs can help to understand the behaviour of these complex computer models and evaluate the robustness of their predictions.

Following expert elicitation of 28 parametric uncertainties, an ensemble of global aerosol models was designed to build emulators of key model outputs. The emulators of the model outputs, such as cloud condensation nuclei and cloud-aerosol forcing, have been used to carry out sensitivity analysis to understand how parameter uncertainties affect the model outputs spatially and temporally and hence understand the behaviour of the model. The sensitivities are then used together with synthetic observations to quantify the value of aerosol observations to constrain aerosol state and aerosol-cloud forcing.

The results have provided aerosol models with unprecedented information on the behaviour of their global aerosol model and have shown how we can exploit the huge database of aerosol measurements. The results also suggest that prediction from an optimal model of aerosol state is unlikely to be a robust prediction of aerosol-cloud forcing.

The statistical methods used were not designed specifically for global aerosol models and are currently being used to gain similar information about other environmental models. Being embedded within the School of Earth and Environment has given me a unique opportunity to apply statistical methods to new problems.

Is it better to accept the offer of a “marginal” liver for transplantation or wait for a potential “optimum” liver?

*Rhiannon Taylor, Elisa Allen, Aaron Goh, Gavin Pettigrew, Dave Collett
NHS Blood & Transplant*

There are more patients awaiting a liver transplant in the UK than there are livers available for transplantation. This has resulted in the increased usage of marginal donor livers and, in particular, organs from deceased donors after circulatory death (DCD). DCD livers carry nearly twice the risk of death post-transplant compared to livers from donors after brain-stem death (DBD). However, the question a patient faces is whether they should accept a DCD liver or remain on the waiting list for a potential DBD liver.

It is logistically and ethically impractical to conduct a randomised control trial to estimate the effect of accepting DCD livers on patient survival compared with remaining on the waiting list for a potential DBD liver. We therefore applied sequential stratification to observational data from the UK Transplant Registry to emulate a hypothetical randomised trial, thus, allowing comparisons to be made between patients who receive a specific treatment with patients waiting for an alternative treatment. Data were used to create strata containing an index case (DCD liver transplant patient) and all patients who were eligible to receive the liver. Data from each constructed stratum were then combined and analysed using a sequentially stratified Cox regression model which estimated separate baseline hazards of mortality for each strata, after adjustment for other relevant risk factors.

The overall results showed that, although DCD livers are associated with worse outcomes following transplantation, the probability of survival is greater if patients accept DCD livers rather than awaiting a potential DBD liver transplant.

Designing, implementing and analysing a virtual trial

*Lauren J Scott, Barney C Reeves, Usha Chakravarthy, Chris A Rogers
University of Bristol*

Background

Neovascular age-related macular degeneration (nAMD) is a common eye condition that can cause severe sight loss and blindness. Active disease is treated monthly until it becomes inactive, but regular monitoring by a hospital ophthalmologist is required as reactivation is common. The ECHoES trial was designed to assess whether, after appropriate training, community optometrists could make decisions about nAMD reactivation, to the same standard as hospital ophthalmologists.

Methods

ECHoES was a non-inferiority virtual trial that utilised an existing repository of images and data collected during the IVAN trial to create 288 patient profiles (vignettes). In a balanced incomplete block design, 96 participants (48 ophthalmologists and 48 optometrists) each reviewed 42 randomly allocated vignettes in a pre-determined order. Each vignette was viewed by 7 ophthalmologists and 7 optometrists.

The primary outcome was correct classification of nAMD reactivation, compared to a reference standard. Data were analysed using mixed model logistic regression, with professional group and vignette-order fitted as fixed effects, and participant and vignette as random effects. The non-inferiority margin was set at 10% assuming ophthalmologists would correctly assess 95% of their vignettes (0.298 on the log-odds scale).

Results

Optometrists and ophthalmologists correctly classified 1702/2016 (84.4%) and 1722/2016 (85.4%) vignettes respectively (odds ratio 0.91, 95% confidence interval 0.66 to 1.25, $p=0.543$).

Conclusion

The ability of optometrists to make nAMD retreatment decisions is non-inferior to that of ophthalmologists. The ECHoES trial was designed in response to a rapid trials funding call and is an example of an efficient trial with a novel design.

Cohorts for studying language development in typically developing populations

*Thomas King, James Law, Cristina McKean
Newcastle University*

Cohorts are now legion and should therefore answer questions of interdisciplinary and cross cultural significance. Language follows a trajectory of natural development in individual children with substantial environmental influences. The socio-cultural variation should be possible to study in cohorts to understand invariants of human language development. Socio-economics needs to be properly controlled, both by using direct measures and sampling a suitable population.

We reviewed the availability of language data in cohort studies around the world by using a mixed search strategy. Systematic database searching for published analyses of data; key informant input; citations in grey literature; and cohort review papers were considered. Inclusion criteria were: two repeated direct measures of language (i.e. expressive or receptive, not phonological awareness) at least one year apart using the same standardised test; a representative sampling scheme for a defined population; peer-reviewed academic publication.

We found fewer than 20 cohort studies meeting these criteria and were able to compose a typology of 5 study types. These are the large demographic cohort; the national education study; the specialised language study; the bio-social development study; and another general epidemiological study. The first two have stronger sampling protocols, better socio-demographic data, larger samples and fewer exclusions but often have limited language measures. Conversely the next two types are smaller and rely on biased sampling such as using recruitment e.g. in ante-natal clinics, or have significant exclusions but tend to have much more sophisticated language measures. Data lack developmental language sophistication and greater interdisciplinary collaboration in design is indicated.

On the robust assessment of cross-over designs against subject drop-out.

*Edward Godolphin, Peter Godolphin
Royal Holloway University of London*

Subject drop-out is a common occurrence in experiments which use cross-over designs, such as clinical trials, where human behaviour or animal response is involved. Typically there are several designs that could be used for an experiment which are equally efficient, however the eventual designs that result from a given drop-out activity may have vastly different properties. In some cases the eventual design is disconnected and then the experiment is damaged severely since little can be achieved from it. To guard against this we establish necessary conditions for an eventual design to be connected, given the loss of potential observation sets due to drop-out, and propose a procedure which assesses possible planned designs for robustness against subject dropout. The method is illustrated by assessing several cross-over designs that have been considered in the recent literature.

Time Series Meets Big Data: Estimating ARMA-GARCH Type Models in High Dimensions

Florian Ziel

European University Viadrina Frankfurt (Oder)

In big data settings, fitting time series models can be a disaster, as common estimation techniques are extremely time demanding. However, lasso estimation methods help to solve many problems with a huge parameter space efficiently.

In this talk we present an iteratively reweighted lasso algorithm for the estimation of time series models under conditional heteroscedasticity in a high-dimensional setting. We analyse the asymptotic behaviour of the resulting estimator that turns out to be significantly better than its homoscedastic counterpart. Moreover, we discuss a special case of this algorithm that is suitable to estimate multivariate AR-ARCH and ARMA-GARCH models in a very fast fashion. Several model extensions like periodic ARMA-GARCH or threshold ARMA-GARCH are discussed as well. We show simulation results that suggest that we benefit more from considering heteroscedasticity in high-dimensional settings than in standard situations. Finally, we present applications to electricity market data.

Novel methods for reptile and amphibian removal data with a hidden state

Ming Zhou, Rachel McCrea, Eleni Matechou, Richard A. Griffiths
University of Kent

Classical removal experiments can be used to estimate the abundance of a population within a closed area, by successively sampling the area for particular animal species and relocating captured individuals to other sites (Zippin, 1956). In the simple case of constant capture probability, removal models result in geometric declines of predicted counts of individuals over time. However, when removals are conducted on reptiles and amphibians, the animals sometimes become undetectable as they hide below ground, and then emerge into the study area from underground when climatic conditions are suitable. Thus we propose a new class of removal models which incorporate a hidden state, allowing the modelling of movement into and out of the observable study area.

The work is motivated by removal data on slow worms, *Anguis fragilis* and common lizards, *Zootoca vivipara*. Removal sampling of these protected species is often undertaken prior to land development. We consider two model developments: (1) a flexible multievent framework (Pradel, 2005) to allow for individuals becoming temporarily unavailable for detection, and (2) accounting for seasonal/climatic variations in detectability and transition probabilities.

We evaluate the performance of the proposed model and consider limitations of our approach by applying it to data from simulations.

Reference

- Pradel, R. (2005) Multievent: an extension of multistate capture-recapture models to uncertain states. *Biometrics*. 61:442-447
- Zippin, C. (1956) An evaluation of the removal method of estimating animal populations. *Biometrics*, 12, 163-189.

Comparison of MCMC Forecasts and Data Augmentation Imputed Data for Item Non-response for Endogenously Allocated Categories

Maria Odejar

Kansas State University

Bayesian Markov chain Monte Carlo (MCMC) algorithms for a supervised learning process are developed to provide estimates and inference about missing data by forecasting without including them in the parameter estimation and straight data augmentation that treats generated data as part of the whole sample. The MCMC algorithms are applied to bivariate ordered threshold model and nested mixed probit model in the context of unresolved cases like don't knows, protests and item non-response in contingent valuation surveys and in sequential discrete decision models respectively. Parameter estimates are compared based on sum of log conditional predictive ordinate goodness-of-fit criterion of the holdout sample data, and predictive ability of the model in and out of the estimation sample based on apparent error rate of misclassification of the estimation data set and the actual error rate of misclassification of the holdout sample. Results show that inclusion of the non-ignorable missing responses thru data augmentation is not detrimental to the quality of parameter estimates, as reflected in their standard deviations, sum of predictive ordinate goodness-of-fit criterion and predictive ability of the model in and out of the estimation sample, as well as in the mixing and convergence properties of the Markov chains generated by the MCMC Bayesian algorithm.

Time series analysis at the Office for National Statistics

James Macey, Duncan Elliott, Jennifer Davies, Cathy Jones, Tariq Aziz, Ping Zong
Office for National Statistics

This poster provides an overview of work carried out by a team of methodologists at the Office for National Statistics working on time series analysis. It provides a brief introduction to time series methods used in the regular production of official statistics, including seasonal adjustment, forecasting and benchmarking and outlines areas of research and collaboration.

Measurement error correction: what method is appropriate?

Christen Gray

London School of Hygiene and Tropical Medicine

Many important exposures in health can only be measured with some amount of error due to the use of imperfect instruments. These include blood pressure, alcohol intake, nutritional values, and income among many others. When measurement error is present in an exposure-outcome relationship, it reduces the power of a study and masks both the shape and magnitude of the true underlying association. There is also evidence that measurement error in national surveys is on the rise; therefore, it is of increasing importance that we have methods available that address it accurately in all settings.

To date, the most prevalent and simple method for measurement error correction is regression calibration. This method is reliable and effective when the underlying relationship is linear, the measurement error is in no way related to the outcome (non-differential), and the odds ratio is not very high. An alternative method is multiple imputation for measurement error correction (MIME). MIME has the advantage of being able to accommodate measurement error which is related to the outcome (differential). The increasingly common use of multiple imputation for missing data also means that the method may be more readily adopted by researchers.

Through the use of simulated data, we compare regression calibration and MIME using validation studies or repeated measures including differential measurement error and common non-linear relationships.

Conflict Contagion

*Tom Clarke, Marie Oldfield
Defence Science and Technology Laboratory (Dstl)*

With an increased emphasis on upstream activity and Defence Engagement, it has become increasingly more important for the UK Ministry of Defence (MOD) and government to understand the relationship between conflict and regional instability. As part of this process, the Historical and Operational Data Analysis Team (HODA) in Defence Science and Technology Laboratory (Dstl) was tasked to look at factors that influenced the regional spread of internal conflicts to help aid the decision making of government.

Conflict contagion is the process by which a conflict in one state (State A) influences the outbreak of conflict in another state (State B). The aim of the task was to produce a tool that could predict the likelihood of conflict contagion based upon a set of numerical variables.

This paper will describe how we conducted this task through a quantitative study of a selection of contagion and non-contagion case studies. We looked at a set of 14 variables that covered structural, political, socio-economic and cultural factors of both States A and B. All case studies took place after the end of the Cold War.

We will discuss the range of statistical methods we used on our dataset, in particular the challenges arising from a dataset containing both binary and continuous data. Progress towards the ultimate aim will also be discussed, together with the outstanding issues that still need to be tackled.

To Combine Forecasts or To Combine Models the Case of Factor and Artificial Neural Network Models

*Ali Babikir, Mustafa Mohammed, Henry Mwambi
Al-Neelain University*

In this paper, the forecasts from the Dynamic Factor Model (DFM) and the Artificial Neural Networks (ANNs) are combined using linear and nonlinear combining methods, the results of the best combining method is compared to the forecasts result of the Factor Augmented Artificial Neural Network (FAANN) hybrid model. The models are applied to forecasts three time series variables using large South African monthly panel. The out-of-sample root mean square error (RMSE) results show that the FAANN model yields substantial improvements over the best combined forecasts from individual forecasting models and the autoregressive AR benchmark model. The Diebold-Mariano test results also further confirm the superiority of the FAANN model forecasts performance over the AR benchmark model and the combined forecasts. The superiority of the FAANN model is due to the ANN flexibility to account for potentially complex nonlinear relationships that are not easily captured by linear models and the role of factors that augmented to the ANN model.

D-optimally constructed design under the variation of non-D-optimality criteria for varying regression polynomials

Polycarp Chigbu, Mary Iwundu
University of Nigeria

The behaviour of D-optimal exact designs, constructed using a combinatorial algorithm, is examined under the variations of A-, E- and G-optimality criteria. In particular, we address the question of whether designs which are optimal with respect to one criterion are also optimal with respect to other criteria. The condition numbers of the designs as well as the equivalence relations of the criteria are established. The D-optimal designs under consideration are for low-order bivariate polynomial models. By the rules of the algorithm, not more than 25% search on the total available designs is required within a design class since a lot of inferior designs, with respect to the optimal design, are eliminated. The models, which could be with or without intercepts, are defined on design regions supported by the points of the circumscribed central composite design. The points are classified into three groups with respect to their distances from the centre of the design region.

Evaluating Positive Behaviour for Learning: a data tetris puzzle.

Rachel Felgate

New Zealand Council for Educational Research

The Positive Behaviour for Learning (PB4L) initiatives aim to help parents, whānau, teachers, early childhood centres and schools address problem behaviour, improve children's wellbeing and increase educational achievement. PB4L is a long-term, systematic approach involving ten initiatives in New Zealand, co-developed with key education sector organisations and is being led by the Ministry of Education. The New Zealand Council for Educational Research (NZCER) was commissioned by the Ministry of Education to evaluate five of these initiatives: PB4L School-Wide, Incredible Years Teacher, Intensive Wraparound Service, Check & Connect, and My FRIENDS Youth. The purpose of these evaluations is to assess how well these initiatives are supporting schools, services, and teachers to provide learning environments or individual support that fosters student engagement and wellbeing.

To gain a representative picture of any of these initiatives multiple data sources are used. Practical considerations mean that these data are not always as easy to link up as ideally wished for. This, alongside the qualitative nature of some data, presents an array of challenges and complications. Some data come from instruments designed specifically for the evaluations, some from Government datasets, and some directly from the schools involved in PB4L.

This paper seeks to examine the complications encountered in this data tetris puzzle and how these multiple data sources provide a coherent picture of PB4L, with a particular emphasis on PB4L School-Wide. A discussion will be given about how these data can be presented to different audiences, and any implications for how the results are utilised.

Non-parametric Regression of Data with Correlated Noise

Tianmiao Wang

University of Bristol

We consider the problem of nonparametric regression with the aim to find a simple approximation that fits given noisy data. The majority of methods in this field have been developed under the assumption that the noise is additive and can be adequately modelled as white noise. However, if the noise is correlated, they do no longer work well. In this poster we focus on the taut string method by Davies and Kovac (2001). We adapt their multi-resolution criterion for data with stationary correlated noise. This requires level-dependent thresholds that depend on the variances of the multi-resolution coefficients on each level and we present an automatic method for estimating these by considering the covariance structure of the noise. In a simulation study and a real data application we compare the performance of the new method with kernel and wavelet based methods for correlated data. The results indicate that our new method not only achieves a smaller mean squared errors on several testbeds, but also provides simpler approximations that better approximate the true features of the observed data.

Statistical Models for estimating the incubation time of AIDS in Kerala state of India

*Brijesh Sathian, Jayadevan Sreedharan, Ritesh G. Menezes
Manipal College of Medical Sciences, Nepal*

Background:

Globally HIV/AIDS is one of the leading challenges for Public Health. The proposed study aims to find suitable statistical distribution for incubation time data on HIV/AIDS.

Methods:

For finding the incubation time distribution, the data of 22 Keralite HIV patients were retrieved from their hospital record through Kasturba Medical College, Mangalore, India. The variables collected were age, gender, incubation time, CD4 cell count, and CD8 cell count. Data were analyzed using EXCELSTAT 2011 and R 2.8.0. A p-value of < 0.01 (two-tailed) for distribution fitting used to establish statistical significance.

Result:

After fitting all applicable distributions of incubation time data of HIV patients, we found that the Weibull (2) distribution fitted best for incubation time of HIV patients; we did parameter estimation. Mean incubation time of Keralite Total (both genders) HIV cases is 4.4, male 5.8 and female 3.4 years.

Conclusion:

It is the first study of the parametric distribution fitting HIV incubation time of Keralite and Weibull 2 distribution appears to offer the best fit. In our study mean incubation time of HIV is 4.4 years; the progression of disease in patients infected with HIV in Kerala seems to be rapid. This is most likely to be due to the late recognition of AIDS-related conditions in the early phase of an epidemic and delayed medical care for patients.

How should we measure residential property prices to inform policy makers?

Jens Mehrhoff

Deutsche Bundesbank

The various motivations for the analysis of house prices, from the monetary policy assessment of price signals to the use for financial stability purposes, or as a soundness indicator, call for alternative measures to be applied. However, these indicators can give different results, which could undermine their credibility for many users. Yet, there should be no unique indicator. The high dimensionality of a complex and diffuse phenomenon such as "the residential property market" cannot adequately be reproduced by a composite indicator. Quite the contrary, the joint distribution of price, financial and real economic indicators seems to be at the centre of the current discussion. There is no simple answer to a complicated question; it might, thus, be better to look at a dashboard of indicators rather than to dissolve existing conflicts between base variables. Last but not least, statistics has a consulting function for policy makers - this makes it even more important to produce unbiased, easily interpretable and manageable measures. In order to determine whether threats to the economy or financial stability emanate from the housing market, the Bundesbank based its analyses on a broad set of indicators. A dashboard comprising the three dimensions price, financial and real economic indicators as well as spatial differences will be discussed. Empirical results for the German residential property market will exemplify the usefulness of a multi-indicator approach in times of strong upward movements of price indicators since 2010; a situation that needs to be addressed in light of the ongoing low-interest-rate environment.

Variational Bayesian inference with Gaussian-mixture trial distributions

Oliver Zobay

MRC Institute of Hearing Research, Nottingham

The Variational Bayes method aims to approximate a complicated posterior p by the minimizer of the Kullback-Leibler distance from a family Q of trial distributions to p . Commonly, Q is chosen as the set of all (completely or partially) factorized distributions. However, the multivariate Gaussian family provides an interesting alternative that enlarges the range of tractable target distributions. This poster presents an extension to the Gaussian approach which uses Gaussian mixtures as approximations.

A general problem for variational inference with mixtures is posed by the calculation of the entropy term in the Kullback-Leibler distance, which becomes analytically intractable. Here, this problem is dealt with by using a simple lower bound for the entropy and imposing restrictions on the form of the Gaussian covariance matrix. In this way, efficient numerical calculations become possible.

To illustrate the method, two- and three-component mixtures are used to approximate an isotropic generalized normal target density and the Bayesian lasso. For heavy-tailed target distributions, the examples show that the mixture approach indeed improves upon single-component approximations in the sense of a reduced Kullback-Leibler distance. From a more practical point of view, mixtures can enhance estimates of posterior marginal variances. Furthermore, they provide an initial estimate of posterior skewness which is not possible with single Gaussians. In addition, general sufficient conditions are derived under which mixtures are guaranteed to provide improvements over single-component approximations.

A comparison of conditional bias-adjusted estimators of a treatment difference for continuous data in group sequential designs

*Masashi Shimura, Masahiko Gosho, Akihiro Hirakawa
Taiho Pharmaceutical Co.*

Background: Group sequential designs (GSD) are sometimes used in clinical trials to determine whether the trial should be stopped early for efficacy or safety concerns. To estimate the mean treatment difference between two treatment groups, a maximum likelihood estimate (MLE) is typically applied as standard; however, it is well known that the MLE is biased. Some researchers have proposed the conditional mean adjusted estimator (CMAE) and the conditional median unbiased estimator (CMUE) as bias-adjusted estimators; however, these estimators are rarely applied in practice because the performance of the estimators has not been adequately investigated.

Methods: We compared the conditional bias of the mean treatment difference and its mean squared error (MSE) among the four estimators (MLE, stagewise MLE, CMAE, and CMUE) through simulation studies in the case of a two-armed clinical trial. Suppose a GSD with one interim analysis for efficacy or futility. Four scenarios was set: (1) the trial stops early for efficacy, (2) does not stop for efficacy, (3) stops for futility, and (4) does not stop for futility.

Results: In Scenarios (1) and (2), the conditional bias and MSE for CMAE were the smallest among the four estimators. On the other hand, CMUE had the smallest bias in Scenario (3). In Scenario (4), the performance of CMAE was similar to that of CMUE and both were better than MLE and stagewise MLE.

Conclusion: CMAE and CMUE are useful to adjust the conditional bias of mean treatment difference in GSD.

A simulation study of spatio-temporal Bayesian models with Poisson response.

Shu Wei Chou Chen

University of Costa Rica

Spatial models that smooth standardized mortality ratios have been used in disease mapping to display and explore spatial patterns due to large variation problems caused by low counts in each area. Recently, analyzing time trends has become of the primary interest in order to display maps throughout a period of time. In spatio-temporal situations where each areal count splits into different time periods, this problem remains because of the presence of more zero counts in each time period and area. Based on annual HIV/AIDS mortality of 81 counties in Costa Rica (1998-2012), alternative spatio-temporal Bayesian models are evaluated using relative bias (RB) and relative standard error (RSE) of the relative risk posterior mean. The relative risks of 81 counties are simulated for a year either by modifying the relative risk of a county or by modifying a cluster of counties in order to produce extreme values. Then, a time series for each county is generated using different time trends. As a result, models show similar spatial smoothing performance based on RB and RSE measures. In addition, it is observed that the temporal performance of these models depends on different time structures. Finally, HIV/AIDS mortality data in Costa Rica (1998-2012) are used as an illustration to compare classic standardized mortality ratios and posterior means of relative risk.

Classification tree models for poverty mapping

*Penelope Bilton, Geoff Jones, Siva Ganesh, Stephen Haslett
Massey University*

The first of the United Nation's Millennium Development Goals is the elimination of extreme poverty by 2015. To achieve this objective, the World Bank has expended billions of dollars of aid in the poorest countries in the world. Optimal distribution of aid resources is facilitated by employing a statistical model to provide estimates, at low geographical levels, of poverty measures, which are then incorporated into a poverty map. ELL, the current standard methodology for poverty mapping, is a small area estimation technique that utilises a linear regression model.

This research applied classification tree models to poverty mapping as an alternative to ELL. The challenges inherent in the research included incorporating complex survey design elements, weighting, stratification and clustering, into tree-based modelling and prediction. Standard errors of prediction for poverty estimates must be obtained indirectly through some type of variance estimation procedure. However, the unstable nature of trees presented major difficulties when building tree models with resampled data.

A designed experiment compared estimate stability and confidence interval coverage using hard versus soft tree estimate types, and jackknife versus bootstrap resampling methods. Soft estimates reduced bias and bootstrapping minimised error. A Monte Carlo study showed that bootstrap soft estimation provided good coverage for data collected by simple random sampling. When cluster effects were introduced into the modelling to simulate data with a complex survey structure, the bootstrap soft estimation procedure still gave reasonable coverage.

Multiple imputation of missing diagnostic codes in a longitudinal healthcare database - a simulation study.

*Aleksandra Turkiewicz, Martin Englund, Jonas Björk
Lund University*

Objectives: To evaluate the multiple imputation (MI) of diagnostic codes in a longitudinal healthcare database in order to estimate the prevalence of a chronic disease in the population.

Methods: We simulated data for 1000 individuals with multiple healthcare visits over time mimicking longitudinal data in a healthcare database. We used a two-level logistic model with covariates on person and visit level to generate the probability of receiving the diagnosis D at the visit. We created 25% or 50% missing diagnoses according to missing completely at random (MCAR) or missing at random (MAR, depending on all covariates) scenario. We created 10 imputed datasets using multivariate normal model for panel data and rounded the imputed values using calibration. We aggregated the data on person level and estimated the prevalence using Rubins rule. We estimated the mean relative bias and mean coverage from 1000 replications of the above procedure for each missingness scenario and a range of prevalences (0.1-0.4) and compared those with naïve estimates (NE) based on the number of persons with the disease registered and the population size.

Results: The MI prevalence estimates had mean relative bias 4.4% (range -13.9% to 19.1%) and mean coverage 0.76 (0.02 to 1.00), similar under MCAR and MAR. NE had mean relative bias of -16.2% (-31.2% to -6.4%) and mean coverage 0.38 (0.00 to 0.96).

Conclusions: The commonly used naïve estimates of prevalence may be considerably biased even if the data are MCAR, because this assumption may not hold on person level. MI was considerably superior but not unbiased.

Survival Analysis of Lung Cancer Patients] {Survival Analysis of Lung Cancer Patients Based on Their Copy Number Alteration Profiles from Next-Generation Sequence Data.

*Khaled Alqahtani, Henry Wood, Pamela Rabbitts, Charles Taylor, Arief Gusnanto
University of Leeds*

Non-small-cell lung cancer (NSCLC) is one of the main sources of death in industrialized nations. As a result, scientists are now looking for some of the risk factors for lung cancer which can be caused by certain changes in the DNA of lung cells. One way to detect these changes is the copy number alteration (CNA); which is a type of structural variation (larger than 1 kbp) in the genome. Like other types of genetic variation, some CNAs have been associated with susceptibility or resistance to disease. As a result, CNA can be used to predict the survival of cancer patients. Next-generation sequencing (NGS) technologies produce high-dimensional data that allow a nearly complete evaluation of genetic variation. With the advent of high-dimensional datasets, the following problem has been faced: the number of covariates (in our study 13968) greatly exceeds the number of observations (85).

The results of our analysis indicate that we can incorporate the copy number alteration profile to predict the survival time. We investigate a Cox proportional hazards model within a random effects model framework using penalized partial likelihood to model the survival time based on lung cancer patients' clinical characteristics as fixed effects and CNA profiles as random effects. We use AIC to estimate Σ which parameterizes the covariance variance matrix of the random effect. For the fixed effects the model indicates that age, stageT3, and stageN2 are statistically significant. Finally, comparing the Kaplan-Meier survival curves with model-based average survival function indicates that the model estimation works reasonably well. Also we covered methods for checking the adequacy of a fitted Cox model.

Simulating positively correlated repeated ordinal data for the proportional-odds model

Nick Parsons

University of Warwick

Ordinal score data are widely reported and analysed across a range of scientific disciplines, often at a number of fixed time-points for each sampling unit in a research study (e.g. from a participant in a clinical trial). These repeated ordinal score data are typically correlated across times for each sampling unit. It is often necessary to simulate repeated ordinal score data to for instance test new analysis methods or determine sample sizes in experimental studies. General methods for simulating correlated ordinal score data are available, but they can be slow and computationally problematic. However, in the specific setting of assumed positive correlation between repeated scores for data from the widely used proportional-odds (PO) model, a particularly simple method can be employed for data simulation. This procedure is described and a simulation study is presented that shows the consistency and reliability of the procedure for a range of typical PO model parameters. One common use for simulating such repeated ordinal scores is in the setting of estimating the required sample size when designing a clinical trial. An example based on the design of the ongoing FIXDT (<http://www.isrctn.com/ISRCTN99771224>) clinical trial comparing methods of fixation after distal tibia fracture is used to demonstrate how the procedure might be used to provide sample size estimates. Implementation is undertaken in R using a simple function that will be made available in the near future.

Modelling Transmission of Vancomycin-Resistant Enterococci in Intensive Care Units

Yinghui Wei, Theodore Kypraios, Phillip O'Neill, Susan Huang, Sheryl Rifas-Shiman, Ben Cooper
University of Plymouth

Vancomycin-resistant enterococci (VRE) are the cause of significant morbidity and mortality among hospital patients. This work is concerned with the analysis of patient-level hospital data on VRE taken from a longitudinal study in a US hospital. The data comprise admission and discharge dates, dates and results of screening tests, and dates during which precautionary measures were in place for each patient during the study period.

We fully describe modelling and inference approaches for various epidemiology of VRE in intensive care units and assess the extent to which infection control measure were effective. We fitted three stochastic models to the patient-level data: 1) Full model, which is a natural generalization of the transmission assumption of the standard SIR model; 2) No-background transmission model, which assumes there is no background transmission; 3) Non-linear model, which assumes that the presence of colonized patients, not their number, affects the rate of new colonisations. All these models are fitted within a Bayesian framework using Markov chain Monte Carlo methods.

Results include estimates of the proportion of unobserved patients colonized with VRE, the proportion of patients colonized on admission, and the efficacy of the control measures within each ward as well as across the eight wards. We compare the three models by using deviance information criteria (DIC). Sensitivity analyses are conducted to assess the robustness of posterior estimates to the choice of priors. Model assessment is conducted by the comparisons between simulated and the observed epidemic data based on the number of colonisations.

Autoregressive transitional ordinal model for the analysis of multivariate ordinal scores in neurology

*Lorenzo Tanadini, Torsten Hothorn, Armin Curt, John Steeves
University of Zurich*

Objectives

For European society, brain-related disorders are a greater socio-economic burden than cancer, cardiovascular diseases and diabetes combined. Nonetheless, pharmaceutical and health companies have been withdrawing from neuroscience because of a large number of failed trials.

A critical appraisal of failed clinical trials in spinal cord injury reveals that a statistically sound analysis of complex ordinal endpoints has yet to be proposed. Past clinical trials with complex ordinal endpoints often resorted to scale collapsing or assumed more refined measurement scales. Both approaches have been shown to be inappropriate in a number of aspects.

Methods and Model

For the analysis of the Upper Extremity Motor Scores in a two-arm RCT, we propose a transitional ordinal model with an autoregressive component.

The proposed model explicitly incorporates the lesion level and the distance from it. The autoregressive component enforces the observed pattern of decreasing neurological scores along the spine of patients.

We investigate the statistical power of the proposed model alongside currently widespread approaches by simulating 1000 clinical trials for different combinations of trial size and treatment effect.

Results

The proposed models had superior statistical power for all simulation settings.

Conclusion

Our approach attempts for the first time in spinal cord injury to take into account the ordinal nature of the endpoint analysed. The analysis is therefore statistically more appropriate and provides clinically better interpretable results. The difference in statistical power is likely to have practical relevance for the planning of future trials.

Statistical Modelling of graphite brick weight loss in advanced gas cooled reactors

*Jenny Burrow, Bryony Hill, Peter Robinson, Philip Maul
Quintessa Limited*

Demonstrating the physical integrity of graphite cores in UK Advanced Gas Cooled Nuclear Reactors operated by EDF Energy (formerly British Energy) is critical to the demonstration of their continued safe operation. The cores contain around 3000 graphite bricks which are subject to cracking and weight loss due to oxidation. Periodic inspections are made of parts of the core that provide information on the state of a sample of the bricks. Statistical models are routinely used to help understand the evolution of brick cracking in the reactors, details of which will be presented in a separate poster. In this poster details are given of the use of mixed effects statistical models for weight loss to predict what will be seen at reactor inspections. By making blind predictions of core behaviour and then comparing these with observations from inspections, the predictive performance of the models can be quantified and can then be used to produce long-term forecasts of core behaviour. A key feature of the models is the need to represent system variability at different scales.

A comparative study of statistical post-processing methods for the calibration of probabilistic weather forecasts

Nina Schuhen, Piers Buchanan, Simon Jackson
UK Met Office

While forecast ensembles allow for the design and usage of novel probabilistic forecast products, they still cannot capture all sources of uncertainty inherent to numerical weather forecasting. In particular they are often not calibrated, resulting in the fact that the probabilistic forecasts derived from ensembles are not statistically consistent with the corresponding observations. A number of statistical post-processing methods for the purpose of calibrating ensemble forecasts have been proposed over the last decade, with Bayesian Model Averaging and Ensemble Model Output Statistics (or Non-homogeneous Gaussian Regression) being among the most successful, as they can be applied to a variety of weather parameters.

At the Met Office, statistical post-processing methods have received more and more attention over the last few years and several calibration techniques based on BMA and EMOS are being trialled and assessed for their benefit over the raw ensemble forecasts. Challenges arise when addressing weather parameters which by nature don't exhibit a normal distribution, like wind speed and precipitation. We will present initial results for the calibration of multiple ensembles, operating on the short- to medium-range, while highlighting the need for preserving the multivariate dependency structure inherent to the ensemble forecasts, both from a site-specific and a gridded point of view.

Comparing surrogate-based uncertainty quantification methods for computationally expensive simulators

Nathan Owen, Peter Challenor, Prathyush Menon
University of Exeter

Computer simulation of physical systems is now ubiquitous in science, with examples including weather forecasting, crash testing and agent-based modelling. The computational models used for this purpose - which we call *simulators* - are typically very complex and can hours, days or weeks to complete just one run. We may make inference about the physical system by making runs of the simulator at various input settings (known as a computer experiment) and performing uncertainty quantification tasks, such as calibration. Traditional approaches to these problems are based around Monte Carlo simulation and involve making a large number of simulator runs. This is simply not possible in the case of expensive and complex simulators, and a more sophisticated approach is to build a surrogate that acts as a fast approximation to the simulator. In this work we compare two popular surrogate methodologies, namely polynomial chaos and Gaussian process emulation. These are contrasting but related approaches to the same problem, but to our knowledge there has yet to be a critical comparison of the two methods in the literature. We use land surface and launch vehicle controller simulators to assess the performance of the two methods in a range of modelling scenarios. Our experiments show that one method does not unanimously outperform the other, but that the preferred method depends on the modelling goals of the practitioner.

Statistical modelling of graphite brick cracking in advanced gas-cooled reactors

*Bryony Hill, Jenny Burrow, Philip Maul, Peter Robinson
Quintessa Ltd*

Demonstrating the physical integrity of graphite cores in UK Advanced Gas Cooled Nuclear Reactors operated by EDF Energy (formerly British Energy) is critical to the demonstration of their continued safe operation. The cores contain around 3000 graphite bricks which are subject to cracking and weight loss due to oxidation. Periodic inspections are made of parts of the core that provide information on the state of a sample of the bricks. Statistical models are fitted to the historic data and used to predict the number of cracked bricks that are expected to be seen at reactor inspections. As more data have become available the complexity of the best performing models has increased. By repeatedly making blind predictions of core behaviour the predictive performance of the models can be quantified and they can then be used to forecast behaviour of the core over longer periods.

The effectiveness of in-vehicle devices at monitoring and reducing levels of driver fatigue

*Simon Tong, Andrew Parkes, Kevin McRae-McKee, Simon Wilson
Transport Research Laboratory*

Driver fatigue is a known concern when considering road safety as it contributes to approximately 20% of all road collisions. Despite the lack of a standardised way to accurately quantify driver fatigue, devices have been developed to monitor fatigue and issue alerts in order to act as a countermeasure; however the performance of such devices requires evaluation.

The aims of this study were to evaluate the performance of these devices, to explore how device alerts affect driver performance and to investigate the effect that fatigue has on fuel consumption. A repeated measures crossover design was implemented using a driving simulator that enabled the collection of multiple variables such as steering angle and lane positioning. Responses were elicited from experts to determine the relative importance of specific variables to calculate a total fatigue score. Having defined a composite fatigue metric, the crossover design allowed for a robust analysis of the effect of fatigue alerts on driver performance and a simple linear regression was initially used to investigate the relationship between fatigue and fuel consumption. As fatigued drivers are expected to behave more erratically than alert drivers, approximate entropy was used as a method to measure the predictability of driver behaviour using the simulator data. While performance statistics provided a better understanding of how well the devices measured fatigue, no significant differences in fatigue scores and fuel consumption were detected between corresponding conditions. The findings from this study have also identified potential areas for further research.

Modeling count data using two different methods: OLS and Poisson Regression

Laura Cascante Amador
Universidad de Costa Rica

Poisson regression is often used for modeling count data. This poster highlights the differences between modeling count data through Ordinary Least Squares (OLS) and Poisson regression to demonstrate how estimated parameters vary from one model to the other. The empirical evidence to illustrate these differences comes from an experiment that evaluated breaking dormancy seed process by counting number of germinated plants. The experiment was adjusted by a factorial design that involved four factors: imbibitions substance, imbibitions time, scarification condition and seed storage temperature before planted. Finally, the experiment integrated twenty treatments including two controls, with three repetitions each.

Due to the nature of the response variable, this kind of data commonly does not distribute normally. Residuals estimated by OLS method were tested to check if normal distribution assumption was met, but results were not conclusive. Kolmogorov-Smirnov test rejected the hypothesis that residuals distributed normally with a probability of 0.033 but, Shappiro-Wilk test confirmed normality with a probability of 0.066. Even when residuals may be considered normal, constant variance was discarded through the corresponding graphic analysis. On the other hand, assumptions for Poisson regression were adequately met, resulting on an under-dispersed model. The analysis showed more significant predictor variables using OLS method than the Poisson regression. Standard errors for every explanatory variable were lower with Poisson regression and, all variables using OLS had the same value for their standard errors. As a conclusion, Poisson regression may provide a more accurate approximation for this kind of data.

Data Visualization for Educational Stakeholders: Enhanced Communication Strategies

Wendy Dickinson

Manatee County Public Schools

This presentation provides unique examples of data display developed to provide statistical information to educational stakeholders. Within United States public school systems, multiple measures of student achievement are generated each academic year. These student achievement measures (typically standardized testing scores) must be analyzed to provide empirical evidence for effective educational decision-making. The empirical evidence must be shared with many audiences: the students and parents, teachers, school administrators, state-level education departments, and the public at large. By utilizing SAS and other software, graphical displays were designed to share student performance on state assessments, and teacher effectiveness as measured by predicted and actual student growth.

Florida is one of several states which requires the utilization of value-added modelling as an indicator of teacher effectiveness. The covariate adjustment model includes covariates such as: the number of subject-relevant courses in which the student was enrolled, up to two prior years of achievement scores, Students with Disabilities (SWD) status, English language learner (ELL) status, gifted status, attendance, mobility (number of transitions between schools), difference from modal age in grade (as an indicator of retention), and similarity (homogeneity) of prior test scores among students in the class.

Teacher effectiveness is reported using four categories: unsatisfactory, needs improvement, effective, or highly effective. Sharing the teacher effectiveness results (classification level) in a visual format helped effectively communicate the information to a wider audience. By using visual display methods we can more effectively inform public consumers of statistical information.

A Regression-Based Investigation into Trends in Coronary Heart Disease and other Cardiovascular Diseases in the UK and Europe since 1980.

Gordon Hunter, Seval Bardak
Kingston University

Heart disease is a major cause of poor health and premature death, particularly in the developed World. However, greater understanding and education about lifestyle-related risk factors, plus improved treatments, have led to mortality and incidence of such diseases declining over recent years. In this paper, we examine trends, both in the U.K. and in various other European countries, in the mortality due to coronary heart disease and prevalence of cardiovascular diseases over time, and the influence of gender and age on these. Various types of regression model are applied. Although most European countries, including the U.K., show an encouraging improvement in these statistics, there are exceptions, and one country shows evidence of poor reporting during the Communist era prior to 1990.

Monetising Data in the Energy and Utility Sectors

*Warren Yabsley, Subramani Venkatachalam, Andrea Gysin, Shirley Coleman
Enzen and Newcastle University*

Data analytics has repeatedly been highlighted as providing a pivotal role in optimising operations and business strategy. Those that embrace this field are shown to gain advantages over organisations that have not adopted such strategies. Enzen Global and ISRU at Newcastle University are engaged in a two-year Knowledge Transfer Partnership (KTP) programme that is part financed by Innovate UK government funding. In the UK, Enzen focuses on Gas, Electricity and Water sectors and works closely with many companies to enhance their practices.

During the KTP project, an enhanced toolbox of data analysis and statistical techniques has been employed to monetise data in a number of areas. For the Gas sector, this has included assessment of pipeline pressure changes and their relationships with time and demand, evaluation and investigation of daily demand changes between different regional distribution networks, and estimation of the effect of variation between local and regional weather on demand predictions. The analysis has been translated into business impact by interpreting the results in monetary terms. For example, in one region, current weather measurement for the region is unrepresentative of local areas contained within the region. This potentially results in over £100,000 per day discrepancy between forecast and actual regional demand.

Due to challenges in the energy sector and the technological improvements in energy distribution, monitoring and control, there is potential for localised information generated through data analytics and statistical analysis. This will influence major investment decisions in energy distribution.

Optimizing route choice in Mini-mountain marathons

Peter Dodd

University of Sheffield

Mini-mountain marathons are a new class of long-format score orienteering event. That is, competitors may choose to visit any checkpoints in any order, but must finish within a fixed time limit. An escalating points deduction schedule severely penalizes late finishes. Checkpoints are given different scores corresponding to their remoteness, and there are more checkpoints on a course than could be visited in the time limit. We analysed data on route choices and splits between checkpoints from the Rab minimountain series, which is available on the web. We obtained checkpoint coordinates from the organisers. We visualised and quantified the variation in route choice. We used linear regression to estimate the relative speed of each competitor and the standardized length of each leg. Using these standardized leg lengths we were able to formulate the competition as an optimization problem and apply a heuristic genetic algorithm to choose routes maximizing score for a given running speed. We found considerable variation in route choice for events without large-scale topographical features. Competitors' score rankings corresponded strongly with their estimated speeds, but with substantial variation beyond the highest 10 finishers. Genetic algorithms were able to select routes which if run at the winner's pace, would have scored more highly.

Preliminary Clinical and Economic Analysis of Diabetes Using a Disease Progression Model

*Shuntaro Yui, Yasutaka Hasegawa, Hajime Sasaki, Takanobu Osaki, Hideyuki Ban, Norman Stein, Sheila McCorkindale, Martin Gibson
Hitachi Europe Ltd*

Currently 10% of the NHS budget is spent treating diabetes and by 2035 this is likely to rise to at least 17% which is clearly unsustainable. Since 2013 Hitachi have collaborated with providers of an integrated diabetes care system in Salford UK and established a cost modelling framework using a Markov model.

Our model was used to validate the degree to which delaying or preventing the onset of type 2 diabetes was economically effective and improved quality of life for patients. Salford has an integrated electronic record system across primary and secondary care (population 242 000) permitting us to perform an analysis across the whole spectrum of care.

We developed a disease progression model using over 9,000 records of patients with pre-diabetes and type 2 diabetes and were able to describe 5 stages of progression. We assigned annual costs to each stage in the model in order to obtain an accurate prediction of future NHS costs.

Using logistic regression analysis fasting blood glucose (or HbA1c) was determined to be the most prominent disease progression factor in the model. We simulated future medical cost savings and found that fasting blood glucose was the best predictor of financial saving (cost difference for the pre-diabetes patients was 2.24 times greater for patients in the lowest category of fasting glucose).

We subsequently quantified the economical effectiveness of using an intervention service to prevent the delaying of diabetes and were able to determine it to be highly cost-effective.

Which measures of cigarette dependence are the strongest predictors of smoking cessation during pregnancy? Analysis of data from a randomised controlled trial

*Muhammad Riaz, Sarah Lewis, Tim Coleman, Robert West, Felix Naughton, Michael Ussher
St. George's, University of London*

OBJECTIVES: To compare the ability of different measures of cigarette dependence to predict smoking cessation in pregnant women

METHODS: This study is based on secondary analysis of data on 784 pregnant women who smoked and were recruited to a randomised controlled trial of physical activity for smoking cessation in pregnancy. We analysed the effects of a number of measures of cigarette dependence, including numbers of cigarettes smoked daily at baseline, Fagerstrom test of cigarette dependence (FTCD) score, and carbon monoxide (CO) level, on smoking abstinence at 4 weeks post-quit and at end-of-pregnancy in logistic regression models. We adjusted each association for age, left full-time education, marital status, self-reported physical activity and use of alcohol. We used adjusted effect sizes and area under the ROC curve to determine which of the predictors perform well.

RESULTS: Lower FTCD score, exhaled CO, and number of cigarettes at baseline were all significant predictors of abstinence at both 4 weeks and end-of-pregnancy. Area under the ROC curve was higher for the models with expired CO and FTCD score (i.e., 0.736 and 0.726, respectively). Heaviness of smoking index HSI, composed of two items of the FTCD, exhibited similar results to the full FTCD.

CONCLUSION: These findings show that the FTCD and HSI scores and exhaled CO may be important predictors of smoking cessation in pregnancy.

Above and Beyond? How Police Officers in England and Wales feel about their employment.

*Deirdre Toher, James Hoggett, Paul Redford, Paul White, Hannah Scott
University of the West of England*

In a major survey of police attitudes towards their work, over 15,500 police officers (over 10% of the full-time equivalent numbers) across 43 police forces in England and Wales completed a substantial questionnaire consisting of over 100 questions about their attitudes to work, along with demographic questions. Response rates varied substantially between forces, as different targeting approaches were used for different forces. However, the results found did not depend on the response rate of the different police forces.

A subset of the resulting data was used to look at the relationship between Organisational Citizenship Behaviour (OCB) and Police Identity (PI). Exploratory Factor Analysis was conducted to measure the primary interests (PI and OCB) and a three-factor solution was found: Police Identity, Organisational Citizenship Behaviour towards Individuals (OCBI) and Organisational Citizenship Behaviour towards the Organisation (OCBO).

Multiple regression was then used to attempt to predict OCB within police officers, where PI is used as one of the predictor variables. The same predictor variables were also used to see whether they predict OCBI and OCBO. After controlling in this manner for other measurable influences, including gender, ethnic group, number of years in service, rank, qualification on entry and geographic region, PI is found to have a significant, positive effect on OCB within police officers. Despite the negativity surrounding the Winsor Review, the associated issues do not appear to affect OCB. The results are consistent with theory and provide an interesting insight into the current nature of policing in England and Wales.

A comparison of correlation matrix modelling methods applied to infant weight Z scores over time

Fraser Tough, Charlotte Wright, John McColl
University of Glasgow

Consider the correlation matrix of n variables X_1, \dots, X_n , a matrix where element ρ_{ij} is the correlation between variables X_i and X_j . These matrices are symmetric with each element of the diagonal equal to one, since the diagonal elements are correlations of the random variables and themselves. Within our research, X_i are infant's weight scores and $i=1, \dots, t$ are time points from 0 to 24 months. It is of interest to model the entries of these matrices so that values can be interpolated at any point on the correlational surface, not just at time points where information is available. The estimates are then taken and used within another application which assesses individual children's weight gain Z scores.

Our research in this area has focused on model comparison - assessing models used within previous research as well as application of some new models. Four modelling approaches were assessed: Generalised Additive Models (GAMs), Fractional Polynomials (FPs), the Cole model (an applied FP model) and the Argyle model (a model based on linear regression). These models were applied to a Malawian dataset, a Pakistani dataset and a South African dataset. A number of different criteria were used to assess the model fit: AIC, BIC, $R^2(\text{adj.})$ and mean square error. The models were applied on different scales, therefore some adjustment had to be made to the AIC and BIC, as models on different scales cannot be compared. FPs and GAMs both fit well, however it is FPs that fit best and are used within further applications.

A sparse hierarchical Bayesian latent variable model for detecting relevant antigenic sites in virus evolution

Vinny Davies, Richard Reeve, William Harvey, Dirk Husmeier
University of Glasgow

Understanding how genetic changes allow emerging virus strains to escape the protection afforded by vaccination is vital both for maintenance of effective vaccines and to understand the antigenic evolution of the virus more broadly. In the current work, we use structural and phylogenetic differences between pairs of virus strains to identify important antigenic sites on the surface of the influenza A(H1N1) virus through the prediction of haemagglutination inhibition (HI) assay, a pairwise measure of the antigenic similarity of virus strains. Previously we have achieved this using a sparse hierarchical Bayesian model for detecting relevant antigenic sites in virus evolution (SABRE). While the SABRE method can account for the experimental variability in the data and select the variables responsible for the changes in the HI assay, it does not fully account for the structure of the dataset. In particular, for any HI assay measurement between the same pair of virus strains the difference in the viral sequence remains the same. Here we propose an extension to the SABRE method (eSABRE) which uses latent variables to represent the underlying HI assay measurement of any given pair of virus strains. This more accurately represents the data in the model while also giving a significant computational improvement. In this work we have applied the eSABRE method to a real H1N1 dataset and identified some of the key antigenic sites on the surface of the H1N1 virus.

Computationally efficient spatial modeling of annual maximum 24-h precipitation on a fine grid

Birgir Hrafnkelsson
University of Iceland

A computationally efficient statistical method is proposed to obtain distributional properties of annual maximum 24-h precipitation on a 1 by 1 km regular grid over Iceland. A covariate based on a local meteorological model that captures information on the physical processes of precipitation is constructed, providing an additional spatial information on maximum precipitation. A latent Gaussian model is built, which takes into account observed maximum precipitation, the covariate based on the local meteorological model, and spatial variations. The observations are assumed to follow the generalized extreme value distribution, where spatial models based on approximate solutions to stochastic partial differential equations are implemented for the location, scale, and shape parameters of the likelihood. An efficient Markov chain Monte Carlo (MCMC) sampler that exploits the sparse matrices induced by the stochastic partial differential equation modeling is implemented, yielding continuous spatial predictions for spatially varying model parameters and quantiles.

A comparison of survival analysis with repeated measurements of a covariate vs. a baseline value only

Munehika Misumi

Radiation Effects Research Foundation, Hiroshima

Identification of predictors of a certain disease risk in the future is one of the major objectives in epidemiological follow-up studies. Some of those studies restrict the investigation to the association of only the baseline value of a variable of interest with the disease risk during follow-up, partly due to the convenience and partly to the popularity of standard Cox regression analyses. However, it is reported that lack of attention to internal time-dependent covariates such as biomarkers could cause attenuation of the association estimate and could affect the conclusion. We applied Cox regression with only a baseline value of a covariate, and compared the result to that obtained by joint models of longitudinal and survival data using repeated measurements of the variable measured with aging of the subjects. We used data of a large cohort of Japanese Atomic bomb survivors followed from 1970 to the end of 2007, and found attenuation of an association estimate of grip strength at baseline to cardiovascular disease mortality when we conducted the usual Cox regression analysis ignoring the repeated measurements of internal covariates. Predictive performance is also compared using both the data only with baseline and the data with repeated measurements. Several joint-modelling approaches have been published in statistical journals. We will show the application of a couple of approaches and results of a simulation study.

Evaluation of parameterization schemes in the Weather and Research Forecasting model through asymmetrical fractional factorial designs

*Fabio Souza, Carla Vivacqua, Michel Mesquita, George Pedra
Bjerknes Centre, Bergen, Norway*

A challenge when using the Weather and Research Forecasting (WRF) model is to identify an appropriate combination of physics options. In this study, we propose the use of asymmetrical fractional factorial designs to investigate the WRF performance for various combinations of parameterization schemes. To illustrate the methodology, we analyze the power of the model in predicting precipitation in the northeast part of Brazil. The objective of the case study is to identify which WRF configurations would reproduce a specific event occurred in the state of Rio Grande do Norte in 2008. The results of the case study indicate that planned experiments are a useful and economical way of evaluating configurations for WRF models. The same approach can be applied to other types of models.

Using Statistical Process Control to explore the utility of the Friends and Family Test

*Sarah Beardon, Alex Bottle, Helen Ward
Imperial College London*

Objectives: A key objective of the Friends and Family Test (FFT) is to highlight variation in patient satisfaction, enabling patient choice and driving improvements in care. We apply Statistical Process Control (SPC) methods to analyse FFT responses to explore patterns of variation and whether unusually poor ratings can be detected.

Methods: Routine FFT data were analysed at both national and local levels, using the proportion of respondents unlikely to recommend the service as the indicator. Both cross-sectional and time-series methods were applied: funnel plots at single time points displayed between-hospital and between-ward variation in ratings. Control charts displayed time trends at hospital and ward level, examining whether unusual changes in patient feedback could be detected.

Results: FFT response rates are generally low, and ward-level data lacked sufficient power to apply SPC methods for analysis, however hospital-level data were more appropriate: for the latest month, 22 out of 448 hospitals in England (4.9%) had higher than expected negative ratings and 2 (0.5%) had lower. Over the last 24 months, hospitals at the local Trust had between one and three time points with unexpected results, however runs which might enable prediction of these occurrences were not present.

Conclusions: There are challenges to using FFT responses for reliable analysis due to lack of power at low level. SPC methods may help to overcome biases in current use of the data, allowing results to be used more sensitively to detect unusual variation. However, obtaining an early warning of poor results may not be possible.

Probabilistic distances between Phylogenetic trees

Maryam Garba, Tom Nye, Richard Boys
Newcastle University

There are many different metrics for measuring how similar two evolutionary trees are, and these are widely used by biologists in a variety of analysis. However, existing metrics ignore that evolutionary trees are really probability models for gene sequence data. We therefore develop metrics between trees based on the properties of the corresponding induced distributions on gene sequence data. We will use the Hellinger distance, the Kullback-Leibler divergence and the Jensen-Shannon distance to measure the distance between evolutionary trees.

Index

A

Abrams, Keith, 101
Adamou, Alexander, 13
Aitkin, Murray, 62
Aldahmani, Saeed, 203
Aldrich, John, 83
Alexander, Craig, 86
Alqahtani, Khaled, 226
Alvarez-Iglesias, Alberto, 53
Anderson, Ben, 23, 77
Ansell, Jake, 63
Artemiou, Andreas, 175
Asar, Ozgur, 110
Ashby, Deborah, 38
Askew, Paul, 17

B

Babikir, Ali, 215
Baillie, Stephen, 162
Bakolis, Ioannis, 94
Banducci, Susan, 31
Beardon, Sarah, 248
Beckley, Nick, 64
Benton, Tom, 18
Bilton, Penelope, 224
Bird, Sheila, 128
Blunt, Gordon, 107
Boehning, Dankmar, 176
Borbely, Tamas, 25
Boyd, Matthew, 129
Bozhiqi, Kristel, 199
Bradford Smith, Kim, 76
Burrow, Jenny, 230

C

Cadman, Emily, 40
Calvert, Carol, 183
Carter, Jackie, 19
Carter, Tess, 56
Cascante Amador, Laura, 235
Cerina, Roberto, 126
Chen, Yining, 123
Chigbu, Polycarp, 216
Chou Chen, Shu Wei, 223
Cole, Diana, 116
Cole, Tim, 171
Cortese, Giuliana, 72
Cortina Borja, Mario, 134
Crane, Eric, 55

D

d. S. Mesquita, Michel, 146
Daniel, Rhian, 20
Das, Sourav, 82
Davies, Gareth, 66
Davies, Vinny, 244

Davison, Anthony, 26
Dawkins, Laura, 60
de Vries, Andrie, 181
Dias, Sofia, 100
Dickinson, Wendy, 236
Dodd, Peter, 239
Dondelinger, Frank, 36
Dray, Matthew, 65

E

Eckley, Idris, 105
Economou, Theodoros, 9
Ejaz, Muslima, 188
Elliott, Duncan, 190
Ellis, Daniel, 68
Elze, Markus, 111
Enki, Doyo, 180
Excell, David, 120

F

Fanshawe, Tom, 49
Farewell, Daniel, 14
Fasli, Maria, 138
Felgate, Rachel, 217
Ferro, Christopher, 122
Fisher, Stephen, 102
Ford, Deborah, 21

G

Galwey, Nicholas, 88
Gandy, Axel, 89
Garba, Maryam, 249
Garratt, Andrew, 127
Gerds, Thomas, 46
Ghosh, Arunava, 197
Gibberd, Alexander, 155
Gittoes, Mark, 45
Gleeson, Emily, 6
Godolphin, Edward, 208
Goldstein, Harvey, 44
Gopal Pillay, Khuneswari, 200
Gray, Christen, 213
Greenaway, Matthew, 143
Guillaumin, Arthur, 16
Gunning, Patricia, 5

H

Haigh, Gregory, 24
Han, Yang, 186
Hand, David, 106, 140
Harris, Richard, 29
Henretty, Nigel, 177
Herbert, Annie, 74
Hewitt, Alan, 78
Hicks, Mike, 166
Hill, Bryony, 233

Hill, Timothy Martyn, 125
Hilliam, Rachel, 50
Hirose, Yoshihiro, 34
Hrafinkelsson, Birgir, 245
Hudson-Smith, Andrew, 98
Hunt, Ian, 167
Hunter, Gordon, 237
Hutton, Jane, 164

J

Jackson, Dan, 35
Jeyam, Anita, 118
Jones, Matthew, 124
Jones, Pete, 7
Jones, Rhys, 30

K

Kadhem, Safaa, 174
Kalktawi, Hadeel, 196
Kejriwal, Mohitosh, 149
Keshinro, Obafemi, 57
King, Thomas, 207
Kontis, Vasilis, 95
Kounali, Daphne, 22
Kunst, Robert, 150

L

Lancaster, Gillian, 48
Langrock, Roland, 133
Lauderdale, Ben, 102
Le Vine, Nataliya, 202
Leckie, George, 43
Lee, Lindsay, 204
Lee, Pete, 201
Lee, Xing, 182

M

Macey, James, 212
Maguire, David, 142
Mann, Hersh, 69
Martin, James, 117
Matechou, Eleni, 132
McCrea, R, 131
McCrink, Lisa, 73
McRae-McKee, Kevin, 234
Mehrhoff, Jens, 220
Mesquita, M, 247
Metcalfe, Liz, 115
Misumi, Munechika, 246
Mitchell, Keith, 195
Morgan, Byron, 141
Mubuuke, Dorothy, 192
Muff, Stefanie, 92
Murr, Andreas, 102

N

Nagaraja, Chaitra, 52

Nardelli, Alberto, 109
Nash, Rachel, 153
Naylor, Jane, 137
Ndagire, Dorothy, 198
Newell, John, 81
Niemi, Aki, 147
Nikolouloupoulos, Aristidis, 136
Noufaily, Angela, 91

O

Odejar, Maria, 211
Olayiwola, Olaniyi Mathew, 169
Oldfield, Marie, 214
Olhede, Sofia, 103, 108
Omar, Rumana, 4
Otekunrin, Oluwaseun, 185
Owen, Nathan, 232

P

Pal, Satyabrata, 184
Park, Tim, 154
Parker, Richard, 67
Parsons, Nick, 227
Paulden, Tim, 119
Payne, Chris, 113
Pearce, Nicky, 114
Pettitt, Tony, 11
Pirani, Monica, 93
Plewis, Ian, 84
Powell, Ben, 158
Puch Solis, Roberto, 163

R

Randell, David, 10
Reeves, David, 130
Rhodes, Kirsty, 151
Riaz, Muhammad, 241
Riley, Richard, 99
Robineau, Delphine, 144
Robinson, Catherine, 42
Rozi, Shafquat, 172
Rozi, Shafquat, 173
Rutterford, Clare, 3

S

Samworth, Richard, 104
Sansom, Philip, 79
Sathian, Brijesh, 219
Sayal, Homesh, 159
Schuhen, Nina, 231
Scott, Lauren, 206
Sei, Tomonari, 32
Shimura, Masashi, 222
Siegert, Stefan, 80
Smith, Ron, 161
Smith, Theresa, 96
Spencer, Neil, 61
Stander, Julian, 135

Stapleton, Matt, 85
Steinacher, Arno, 152
Stephenson, David, 59
Strong, Colin, 139

T

Takeda, Jun, 187
Takemura, Akimichi, 33
Tampubolon, Gindo, 12
Tanadini, Lorenzo, 229
Tawn, Jonathan, 27
Taylor, Rhiannon, 112, 205
Thorarinsdottir, Thordis, 47
Toher, Deirdre, 242
Tompsett, Daniel, 178
Tough, Fraser, 243
Tudur Smith, Catrin, 41
Turkiewicz, Aleksandra, 225

V

van den Brakel, Jan, 157
Vanhatalo, Jarno, 145
Ventrucci, Massimo, 97
Vickers, Lucy, 39

W

Walker, Sarah, 179
Wallbank, Caroline, 193, 194
Walshaw, David, 28

Walter, Stephen, 54
Wang, Dandan, 191
Wang, Tengyao, 156
Wang, Tianmiao, 218
Waruszynski, Paul, 75
Wei, Yinghui, 228
White, Chris, 70
Whittaker, John, 37
Wild, Chris, 51
Williams, Andrew, 189
Wilson, Amy, 90, 165
Wyszynski, Karol, 15

Y

Yabe, Ryota, 148
Yabsley, Warren, 238
Yaya, OlaOluwa, 170
Yildiz, Dilek, 8
Youngman, Ben, 58
Yui, Shuntaro, 240

Z

Zeger, Scott, 71
Zelin, Andrew, 121
Zhou, Ming, 210
Zidek, James, 160
Ziel, Florian, 209
Zobay, Oliver, 221

The logo for the Royal Statistical Society, featuring a vertical yellow bar to the left of the text "ROYAL STATISTICAL SOCIETY".

ROYAL
STATISTICAL
SOCIETY

DATA | EVIDENCE | DECISIONS

RSS International Conference • 7 – 10 September 2015 • Exeter University
www.rss.org.uk/conference2015