ROYAL
STATISTICAL
SOCIETY
DATA | EVIDENCE | DECISIONS

RSS International Conference

Abstracts booklet

GLASGOW

4 – 7 September 2017

www.rss.org.uk/conference2017

Abstracts are ordered in date, time and session order

Plenary 1 - 36th Fisher Memorial Lecture
Monday 4 September – 6pm-7pm

And thereby hangs a tail: the strange history of P-values

Stephen Senn
Head of the Competence Center for Methodology and Statistics, Luxembourg Institute of Health

RA Fisher is usually given the honour and now (increasingly) the shame of having invented P-values. A common theme of criticism is that P-values give significance too easily and that their cult is responsible for a replication crisis enveloping science.

However, tail area probabilities were calculated long before Fisher and naively and commonly given an inverse interpretation. It was Fisher who pointed out that this interpretation was unsafe and who stressed an alternative and more conservative one, although even this can be found in earlier writings such as those of Karl Pearson.

I shall consider the history of P-values and inverse probability from Bayes to Fisher, passing by Laplace, Pearson, Student, Broad and Jeffreys, and show that the problem is not so much an incompatibility of frequentist and Bayesian inference, as an incompatibility of two approaches to dealing with null hypotheses. Both these approaches are encountered in Bayesian analyses, with the older of the two much more common. They lead to very different inferences in the Bayesian framework but much lesser differences in the frequentist one. I conclude that the real problem is that there is an unresolved difference of practice in Bayesian approaches. This does not, of itself, constitute a defence of P-values but it does suggest that some of the problems for which they are blamed will not be resolved merely by abandoning them.

1.1 Contributed – Medical Statistics: Clinical Trials
Tuesday 5 September – 9am-10am

The analysis of two-stage randomised trials when some participants are indifferent to treatment choice

Stephen Walter, Robin Turner, Petra Macaskill, Kirsten McCaffery, Les Irwig
McMaster University

Participants in clinical trials would often prefer to receive one of the treatments being compared, if that were allowed. Effects of preferences on study outcomes can be important, but they are unidentifiable in standard trial designs. However they can be estimated using a two-stage trial design, in which a random subset of patients are indeed permitted to choose their treatment, while the remaining patients are randomised in the usual way.

We have previously discussed the optimum design and analysis of the two-stage randomised trial, but we assumed that all participants have a preferred treatment [1,2]. We now extend this work to allow for some participants to have no preference, even if they are offered a choice.

We obtain unbiased estimates and tests of the effects of preferences on study outcomes, as well as the usual direct treatment effect, in this context. This approach will be illustrated using a medical *versus* surgical trial in which 69% of participants in the choice arm had no preferred treatment.

We conclude that the two-stage design remains attractive even if some participants have no stated treatment preference. It allows insight into potentially important determinants of study outcomes that cannot be estimated in other trials. Furthermore the outcomes in trials participants with various treatment preferences, including being indifferent, are themselves of interest to trial investigators, and may inform optimum treatment decision-making for future patients.

1. Walter SD, Turner RM, Macaskill P, McCaffery KJ, Irwig L. Beyond the treatment effect: evaluating the effects of patient preferences in randomised trials. Statistical Methods in Medical Research, 2014.
2. Walter SD, Turner RM, Macaskill P, McCaffery KJ, Irwig L. Optimal allocation of participants for the estimation of selection, preference and treatment effects in the two-stage randomised trial design. Statistics in Medicine 31, 1307-22, 2012

1.1 Contributed – Medical Statistics: Clinical Trials
Tuesday 5 September – 9am-10am

Covariate adjustment and prediction of mean response in randomised trials

Jonathan Bartlett
AstraZeneca

A key quantity which is almost always reported from a randomised trial is the mean outcome in each treatment group. When baseline covariates are collected, these can be used to adjust these means to account for imbalance in the baseline covariates between groups, thereby resulting in a more precise estimate. Qu and Luo (DOI: 10.1002/pst.1658) recently described an approach for estimating baseline adjusted treatment group means which, when the outcome model is non-linear (e.g. logistic regression), is more appropriate than the conventional approach which predicts the mean outcome for each treatment group, setting the baseline covariates to their mean values.

In this talk I will first emphasize that when the outcome model is non-linear, the aforementioned `conventional' approach estimates a different quantity than the unadjusted group means and `Qu and Luo' estimator. I will then describe how for many commonly used outcome model types, the Qu and Luo estimates are unbiased even when the outcome model is misspecified. Qu and Luo described how standard errors and confidence intervals can be calculated for these estimates, but treated the baseline covariates as fixed constants. When, as is usually the case in trials, the baseline covariates of patients would not be fixed in repeated sampling, I show that these standard errors are too small. I will describe a simple modification to their approach which provides valid standard errors and confidence intervals.

I will then discuss the impact of stratified randomisation and missing outcomes on the preceding results, and give suggestions for when baseline adjusted means may or may not be preferred to unadjusted means. The analytical results will be illustrated through simulations and application to a recently conducted trial with recurrent events analysed using negative binomial regression.

1.1 Contributed – Medical Statistics: Clinical Trials
Tuesday 5 September – 9am-10am

Common pitfalls in Oncology Trials

Arnaud Demange, Alan Phillips, Jia Ma
ICON Clinical Research

Oncology trials fail far too frequently, and often for similar reasons. Understanding the risks in oncology trial designs and analyses, and how to mitigate them, will equip developers to reduce the failure rate in late stage oncology trials and move effective drugs to market faster.

We have observed oncology studies fail frequently due to five prevalent issues:

1. assumptions that undermined a trial’s design;
2. inadequate patient selection criteria;
3. problematic phase 2 trials;
4. over-interpretation of subgroup analyses; and
5. optimistic selection of a non-inferiority margin.

This presentation will review these common issues in the context of specific failed trials and propose practical countermeasures that would not compromise study integrity, including adopting an adaptive design and deploying a more stringent statistical significance level to minimize the false discovery rate.

1.2 Contributed - Official statistics and public policy: **Use of the Census**
Tuesday 5 September – 9am-10am

An online census: Will people respond online? Characteristics of online and paper respondents

David Corps, Victor Meirinhos
Office for National Statistics

The 2021 Census in England and Wales will be a primarily online census for the first time. To understand further the willingness to respond online and the characteristics of those who may be unwilling or unable to respond online, a large-scale test was designed for early 2017. Sixty thousand households were sent either a paper questionnaire or a letter including a unique access code enabling online response. Anyone receiving a unique access code could request a paper questionnaire and those who received a paper questionnaire could choose to respond online.

Here, we discuss the differences in response rates by mode for those given a paper questionnaire and those sent a unique access code. Furthermore, we present the analyses of characteristics of respondents, using a mixed effects modelling approach to investigate differences in characteristics among those who respond via paper or online for those given a paper questionnaire, and those responding online or via paper for those sent a unique access code.

This research will play an important role in informing the design of the 2021 Census and in understanding more about the digitally excluded population in general in an increasingly digital world.

1.2 Contributed - Official statistics and public policy: **Use of the Census**
Tuesday 5 September – 9am-10am

New opportunities and challenges in census coverage assessment

Viktor Racinskij
Office for National Statistics

By definition census is a complete enumeration of a population of interest. However, when dealing with large and complex human populations census coverage errors are inevitable. Estimation of the coverage errors has been a vibrant area in statistics for over sixty years. In the context of the Census of England and Wales, the work related to the coverage error estimation is known as coverage assessment and it is one of the key components in producing high quality population statistics.

Innovations in data collection and pre-processing open a wider range of opportunities for the coverage assessment methods compared to what was available in the past. Unsurprisingly, the new opportunities bring new challenges. This talk covers the most recent research undertaken by the Office for National Statistics in the area of census coverage assessment with a view to reconcile new opportunities and challenges.

Central to the research is modelling of the linked 2011 Census and Census Coverage Survey data using multilevel logistic models. A number of models are explored with a particular interest in modelling age effect with the alternatives ranging from the simple quinary group effects to a continuous effect modelled using restricted cubic splines. These models are further employed in simulation studies to assess performance of estimation approaches such as the dual system estimation carried out at the age-sex by cluster of small areas stratum, deeply post-stratified dual system estimation and logistic regression estimation.

The outcomes of competing assessment approaches are analysed and compared with the goal of finding the approach that simultaneously uses the advantages of the new opportunities and prevails over the corresponding challenges in a reasonable manner.

1.2 Contributed - Official statistics and public policy: **Use of the Census**
Tuesday 5 September – 9am-10am

An innovative approach to estimating the costs of the 2021 Census Coverage Survey

Adriana Castaldo
Office for National Statistics

The Census Coverage Survey (CCS) is a post-enumeration survey that is conducted shortly after the census in order to assess the coverage of the census. This paper describes the research that the Office for National Statistics (ONS) is doing to assess the impact of using different sampling design options on the costs of running the CCS. The focus of the research is on investigating the impact of reducing the level of clustering in the CCS sample on the associated costs of fieldwork. It is well known that by reducing the level of clustering in a sample survey there is a gain in the statistical efficiency of the associated estimates, but potentially at increased costs, due to greater travel time for the interviewers to reach more scattered addresses. The two key costs that are affected by the sampling design are the number of interviewers and travelling costs. These are being evaluated using a variation of a cost model for clustered surveys that was developed for the design of the 2011 CCS. One key new feature in the current work compared to 2011 is the use of an automated zone design tool developed by the University of Southampton, to estimate interviewers' workload allocation of the CCS sample. The use of this tool integrated within an appropriate cost model for clustered surveys will allow for more efficient and realistic calculations of the costs of running the CCS. The paper will present the results of this new approach to calculating the costs of the CCS and has potential for wider applications within and outside ONS.

1.3 Contributed - Environmental/Spatial statistics: **Environment and Health**
Tuesday 5 September – 9am-10am

Air Health Indicator: Does ozone's adverse health effect depend on age, sex and region?

Hwashin Hyun Shin, Mamun Mahmud, Wesley Burr, Marc Smith-Doiron, Branka Jovic, John Than
Health Canada

Introduction: The Air Health Indicator (AHI) is designed to estimate Canadian population health risk related to **short-term exposure** to air pollution and to detect trends in the national annual health risks. Potential differences in the national health risks by age, sex and region have been in question.

Objective: The goal is to estimate and find trends in cardiopulmonary (CP) mortality risks related to short-term exposure to ground-ozone (ozone) at national and regional levels for **warm season** (April-Sept.) between 1984 and 2012.

Methods: 24 major cities were assigned to 3 regions: 5 cities to East, 13 cities to Centre and 6 cities to West. An estimator using 7-year blocks was employed to find trends in annual associations through a Bayesian hierarchical model. For age grouping, all ages 1 and above have been classified into three age groups: A0 (≥1), A7(> 50) & A8 (>65).

Results: The baseline national risk from all 29 years combined is 0.0008, which means a 0.8% increase in CP mortality is associated with a 10 ppb increase in daily Ozone, with a 95% credible interval of (0.0003, 0.0012). While age group A7 returned the same risk estimates, age group A8 showed slightly higher estimates, (0.0004, 0.0013). On the other hand, females seem to be more vulnerable than males: (0.0005, 0.0017) vs (-0.0003, 0.0010). Among the regions, the central region showed higher ozone risk (0.0001, 0.0009). From the 7-year estimator, a non-linear trend was found, demonstrating a mixed trend with an incline, a flat, and a decline regardless of age and sex. The between-city heterogeneity showed an overall decline.

Discussion: The findings quantified the differences in ozone's effect by age and sex. While ozone's effects on CP mortality have declined for later years, CP-related prescribed medications have increased significantly since 2001, the link of which needs further investigation.

1.3 Contributed - Environmental/Spatial statistics: Environment and Health
Tuesday 5 September – 9am-10am

Using Low Cost Sensors to Measure Air Pollution

Yoana Borisova, Marian Scott, Duncan Lee
University of Glasgow

Air pollution in the United Kingdom is measured using monitors which are accurate but expensive. The Centre for Ecology and Hydrology (CEH) operates such a network of monitors. This paper reports on the performance of a type of lower cost sensors. Data were collected from two identical sensors placed next to a CEH monitor just outside Edinburgh, and nitrogen dioxide, NO_2, and ozone, O_3, measurements were taken at 15 minute intervals. Two types of emission voltages (labelled AE and WE) were recorded for each of the pollutants. Additionally, the temperature and relative humidity were also recorded.

The two low cost sensors are compared to each other using correlation and Bland-Altman analysis. It is found that for NO_2 both types of emission voltages, AE and WE, taken by the two sensors are significantly different from each other. For AE, the two sensors had, on average, a difference of about 100 volts, whereas for WE, the two sensors had, on average, a difference of about 200 volts. However, for O_3, WE was shown to be on average, the same in both sensors.

Finally, the performance of the low-cost sensors is compared to the measurements taken by the CEH monitor. Using linear regression and generalised least squares models, it was found that the SE voltage is statistically related to the meteorological conditions but not the reference pollutant concentrations. However, for the WE voltage, there is a significant relationship between the voltages and the pollutant concentrations in addition to the meteorological conditions. The best models have $R^2_{adj.}$ of about 50% indicating that half the variability in the data can be explained by the models. Therefore, it is concluded that the low-cost sensors (WE voltage) show promising results in measuring the NO_2 and O_3 concentrations.

1.3 Contributed - Environmental/Spatial statistics: **Environment and Health**
Tuesday 5 September – 9am-10am

Global Estimation of Air Quality and the Burden of Disease associated with Ambient Air Pollution

Gavin Shaddick
University of Exeter

Air pollution is a major risk factor for global health, with both ambient and household air pollution contributing substantial components of the overall global disease burden. One of the key drivers of adverse health effects is fine particulate matter ambient pollution (PM2.5) to which an estimated 3 million deaths can be attributed annually. The primary source of information for estimating exposures has been measurements from ground monitoring networks but, although coverage is increasing, there remain regions in which monitoring is limited. Ground monitoring data therefore needs to be supplemented with information from other sources, such as satellite retrievals of aerosol optical depth and chemical transport models. A hierarchical modelling approach for integrating data from multiple sources is proposed allowing spatially-varying relationships between ground measurements and other factors that estimate air quality. Set within a Bayesian framework, the resulting Data Integration Model for Air Quality (DIMAQ) is used to estimate exposures, together with associated measures of uncertainty, on a high-resolution grid covering the entire world. Bayesian analysis on this scale can be computationally challenging and here approximate Bayesian inference is performed using Integrated Nested Laplace Approximations. Based on summaries of the posterior distributions for each grid cell, it is estimated that 92% of the world's population reside in areas exceeding the World Health Organization's Air Quality Guidelines. Estimated exposures from the model, produced on a high-resolution grid (10km x 10km) covering the entire globe, are combined with risk estimates to produce a global assessment of exposures to PM2.5 and to estimate the associated burden of disease attributable to air pollution.

1.4 Contributed - Social statistics: Methodology
Tuesday 5 September – 9am-10am

Lexis surface visualisations: a vitally important tool for seeing and understanding patterns in population data

Jonathan Minton
University of Glasgow

A Lexis surface is the result of visually arranging a population-level variable z - such as population exposure, mortality risk, or fertility - on a surface defined by relative time (usually age) on one axis and absolute time (usually year) on another axis. Conceptually, Lexis surfaces are as to time as topographic maps are to space, substituting latitude for age, longitude for year, and height above sea level for the variable z. Cartography is the long-established science and art of representing a three dimensional surface effectively on a two-dimensional canvas, and so many of the techniques developed in cartography for representing spatial surfaces - such as the use of contour lines to indicate positions of equal 'height' on a surface - can also be usefully applied to the visualisation of temporal population trends in Lexis surfaces. Advancements in interactive 3D data visualisation and computer-aided manufacture mean that Lexis surfaces can 'escape the canvas' and be explored as interactive virtual three dimensional 'objects', or even produced using 3D printers and explored and interacted with directly, turning data exploration into a tactile as well as visual experience.

This talk will comprise three parts. In part one I will discuss my introduction and initial applications of Lexis surfaces, along with their longer term history and precedents. In part two I will describe a number of variants of Lexis surface visualisations I have developed and seen developed by other researchers, the substantive problems they have been developed to address and some of the key findings developed through their application. Finally, in part three I will discuss the future of Lexis surface visualisations, and make the argument that they may have a key role to play in the development of more substantively sophisticated statistical models of population processes, as well as in the effective visual communication of population data.

1.4 Contributed - Social statistics: Methodology
Tuesday 5 September – 9am-10am

Investigating and Optimising Patients’ Medicines Networks after Discharge from Hospital: A Multilevel Approach.

Mark Tranmer
University of Glasgow

Patients are at heightened risk of harm from their medicines when their care is transferred between providers, often because of poor information sharing and limited opportunities to discuss managing their medicines with healthcare professionals. Patients may have a range of medicines contacts influencing how their discharge medicines are used. This research explores the structure of patients’ personal and professional medicines networks after they are discharged from hospital, and the functions provided by those networks. The study was conducted with 61 patients discharged from the cardiology wards of two acute hospital trusts. Interviews were conducted six weeks after hospital discharge. Participants’ ages ranged from 35-80; 43 were male and 18 female. Patients had between 1 and 15 individual medicines contacts in their networks to whom they attributed a score on various aspects of medicines support. In each case these contacts can be conceptualised as a tie between the patient and a member of the patient's support network. Multilevel models were applied to these data, where the unit of analysis is the tie between the ego (the patient) at level two and the alter (the member of the support network of the patient) at level one. The multilevel model framework allows an investigation of variation in the ego-alter ties, given the specific nature of the tie (e.g. support for condition management, or for orientation) and the way in which variations in the tie values are associated with characteristics of ego and alter. E.g. gender of patient, gender of support person, and whether the support person is a family member or a medical professional. The multilevel model framework thus allows a range of substantive research questions to be answered in this context.

1.4 Contributed - Social statistics: Methodology
Tuesday 5 September – 9am-10am

An extension of the general 3-step ML approach to random effects event history models with multiple latent categorical predictors

Yajing Zhu, Fiona Steele, Irini Moustaki
London School of Economics and Political Science

It is common in social research for more than one predictor variable to be a latent construct and there are many applications of structural equation models (SEM) with multiple continuous latent variables as predictors of one or more distal outcome. Researchers may wish to treat these latent constructs as categorical, but until recently have been limited to methods such as the modal class approach, which does not account for misclassification, or the 1-step approach, which creates a circular relationship in which the latent variable is also partly measured by the outcome. In this talk, we discuss an extension of the 3-step approach for one latent variable to a random effects event history model for recurrent events where the hazard function depends on multiple associated latent categorical variables. We describe maximum likelihood estimation of such a model and its potential to generalise to more flexible structural equation models that can handle longitudinal and other forms of clustering, measurement error and mixed response types.

1.5 Contributed - Methods and theory: Time Series
Tuesday 5 September – 9am-10am

Optimal bias-correction in the log periodogram estimation of the fractional parameter: A jackknife approach

Kanchana Nadarajah, Gael Margeret Martin, Don Poskitt
Monash University, Australia

In this paper, we develop a new bias-corrected log-periodogram regression (LPR) estimator of the fractional parameter, d, in a stationary fractionally integrated model, by using a jackknife technique. The optimal weights of the full sample and the sub-samples used to construct the jackknife estimator are chosen such that bias reduction occurs to an order of $n^{-\alpha}$ for some $0<\alpha<1$, without the usual increase in asymptotic variance being incurred. The optimal weights involve two types of covariance terms related to the log periodogram: (i) covariances between the full sample and the sub-samples, and, (ii) covariances across sub-samples. These covariance terms may also be represented by cumulants of the discrete Fourier transform (DFT) of the time series. We show that DFTs at different frequency ordinates are asymptotically independent. We then obtain a general formula for the covariance terms related to the periodograms, and show that the periodograms are asymptotically independent chi-square random variables. Thereby, a general expression for the covariance terms in (i) and (ii) are also obtained. Assuming that the number of sub-samples is fixed, we show that under some regularity conditions our jackknife estimator is consistent and the limiting distribution is normal with the same asymptotic variance and rate of convergence, $n^{\alpha/2}$, as the original LPR estimator. These theoretical results are valid under both non-overlapping and moving-block sub-sampling schemes used in the jackknife technique. A Monte Carlo study shows that the optimal jackknife estimator outperforms the other bias-reduced estimators - namely the pre-filtered sieve bootstrap-based estimator of Poskitt, Martin and Grose (2016) and the weighted average estimator of Guggenberger and Sun (2006) - in terms of bias-reduction and root-mean-squared-error. The optimal jackknife estimator is then applied to two empirical time series: (i) the annual minimum level of the Nile river, and (ii) realized volatility for the S&P500 stock index.

1.5 Contributed - Methods and theory: Time Series
Tuesday 5 September – 9am-10am

Simulation-based selection of prediction models

Robert Kunst
Institute for Advanced Studies

This work assesses the benefits of basing model selection decisions in a forecasting context on simulations that fuse data information and the structure hypothesized by tentative rival models. These procedures can be applied to any empirical forecasting problems. In the limits of this project, however, the focus is on macro-economic applications.

The suggested procedure aims at choosing among a small number of tentative forecast models in the presence of data. From models fitted to the data, pseudo-data are generated. Again, the models are applied to the pseudo-data and their out-of-sample performance is evaluated. The ultimate choice of the forecasting model is based on the relative performance of rival models in predicting "their own data" and those of the rival model.

The project covers the three aspects of a rigorous statistical foundation, of a Monte Carlo evaluation of the procedure, and of exemplary empirical implementations of the method in typical macro-economic applications.

1.5 Contributed - Methods and theory: Time Series
Tuesday 5 September – 9am-10am

Verifying predictions

Mark Davis
Imperial College

This talk concerns the computation of predictions, such as risk measures for financial data, and asks how, given a forecasting procedure, we can tell whether the answers it produces are correct. We concentrate on situations where we observe data in real time, make predictions and observe outcomes. It is argued that evaluation is best addressed from the point of view of probability forecasting or Dawid's theory of `prequential statistics' [Dawid, JRSS(A)1984]. We introduce a concept of `consistency' of a risk measure, which is close to Dawid's `strong prequential principle', and examine its application to quantile forecasting (for example, Value-at-Risk (VaR) in finance). It turns out that consistency is essentially a dynamic version of `elicitability' [Gneiting, JASA 2011]. We show in particular that VaR has special properties not shared by any other risk measure. We show that a simple data-driven feedback algorithm can produce VaR estimates on financial data that easily pass both the consistency test and a further newly-introduced statistical test for independence of a binary sequence. The results are equally applicable in other areas such as weather forecasting.

1.6 Contributed - Communicating and teaching statistics: Statistical literacy and training in schools and the workplace
Tuesday 5 September – 9am-10am

Student active learning online and in the classroom by combining the best of Flipped Classroom and MOOCs when teaching statistics

Maria Karlsson, Anders Lundquist, Mathias Lundin
Umeå University

Flipped classroom teaching means, in short, to change/flip where students and teachers do things. Instead of students (passively) listening to a teacher's lecture in the classroom and then going home to solve problems on their own, the idea is that students first watch video lectures at home and then come to the classroom to (actively) solve problems together with other students and the teacher.

As part of a pedagogical project, we have used special learning management system (LMS) called "Scalable Learning", which facilitates teaching with flipped classroom techniques. The Scalable Learning LMS allows the teacher to make the video lectures interactive by adding quizzes and surveys to the videos (as is very common in Massive Open Online Courses (MOOCs)), thereby helping the students to be more active while watching the videos. Moreover, there are different tools implemented in the LMS so that students e.g., can ask questions in direct connection to the videos. The student supplied questions and their answers to the quizzes can then, together with other data from the LMS, be used by the teacher to prepare classroom activities customized to the students' needs.

We present our experiences of teaching a course in statistics using this LMS; there are both pros and cons from a teachers' point of view. Also, we present the students' opinions with the LMS and with the flipped classroom teaching that they were exposed to during the course. Most of them were very satisfied with the concept but they also provided us with several interesting ideas on how to improve the course.

1.6 Contributed - Communicating and teaching statistics: Statistical literacy and training in schools and the workplace
Tuesday 5 September – 9am-10am

Getting down and dirty with data: how Q-Step Centres are training students in the workplace

Jackie Carter
University of Manchester

The presentation will take the form of a panel session. Carter will set the context (of Q-Step as a national initiative to improve statistical literacy in the social sciences graduate population) and then introduce the 4 student speakers, from the Universities of Manchester, Edinburgh, Queens Belfast and Cardiff, to represent the 4 nations of the UK. Each will speak for 3 minutes only - using the 3-minute-thesis format adapted from the University of Queensland. Each will represent a different social science discipline (sociology, politics and international relations, criminology, social analytics). Each represents a different work placement model.

In the remaining 5 minutes Carter will present the findings of the Q-Step programme in support of applied research skills, as outlined in her book. Examples of student outputs (blog posts, reports, academic papers, dissertations) and follow-on work (opportunities at doctoral level with the organizations involved at undergraduate level) will be presented.

The aim of the session is to provide the audience with a variety of experiences through the Q-Step centres, demonstrating how statistics teaching can be consolidated through experiential learning. The work-place hosts are from the public, private and voluntary sector, and include government departments and international organizations.

The Q-Step programme is experimental, designed to introduce a step-change in teaching statistics to social science undergraduates. Lessons learned from the programme to date (since Oct 2013), will be shared - with a focus on the practical nature of learning data analysis. Ideas for expanding this model to other subject areas will be presented, in brief, with some examples of how Q-Step placements have already provided opportunities to the UK doctoral training centres.

Finally, the Q&A session will invite comments on how we can better develop a virtuous circle of university statistics education in the social sciences, alongside employers' needs.

1.7 Contributed - Data science: Clusters and networks
Tuesday 5 September – 9am-10am

Extracting Meaningful Patterns from Big Binary Data using E-BiBit Algorithm in R

Ewoud De Troyer, Ziv Shkedy, Adetayo Kasim
Hasselt University (CenStat)

Biclustering is a data analysis method that can be used to cluster the rows and columns in a (big) data matrix simultaneously in order to identify local submatrices of interest, i.e., local patterns in a big data matrix. For binary data matrices, the local submatrices that biclustering methods can identify consists of rectangles of 1's. Several methods were developed for biclustering of binary data, such as the *Bimax* algorithm proposed by Prelić et al. (2006) and the *BiBit* algorithm by Rodriguez-Baena, Perez-Pulido, and Aguilar-Ruiz (2011). However, these methods are capable to discover only perfect biclusters which means that noise is not allowed (i.e., zeros are not included in the bicluster). We present an extension for the *BiBit* algorithm (*E-BiBit*) that allows for noisy biclusters. While this method works very fast, its downside is that it often produces a large number of biclusters (typically >10000) which makes it very difficult to recover any meaningful patterns and to interpret the results. Furthermore many of these biclusters are highly overlapping.
We propose a data analysis workflow to extract meaningful noisy biclusters from binary data using an extended and `pattern-guided' version of BiBit and combine it with traditional clustering/networking methods. The proposed algorithm and the data analysis workflow are illustrated using the **BiBitR** R package to extract and visualize these results.

The *E-BiBit* has also been included in the **BiclustGUI** R package, an ensemble GUI package in which multiple biclustering and visualisation methods are implemented.

1.7 Contributed - Data science: Clusters and networks
Tuesday 5 September – 9am-10am

A soft Adjusted Rand Index for comparing probabilistic clusterings

Nema Dean, Abby Flynt
University of Glasgow

The adjusted Rand Index (ARI) is a metric used to compare two clusterings of the same dataset. If the results of the two cluster analyses are similar, the ARI value will be close to 1. The theoretical properties of the index are well understood and it is fairly intuitive. However, it assumes that the result of each cluster analysis is a partition which assigns each object to a single cluster. This ignores the information that can come from such techniques as model-based clustering which produce such assignments on the basis of the cluster membership probabilities estimated from the model. These probabilities have the advantage of still identifying the most likely cluster but also giving a measure of uncertainty in that assignment. Since the ARI only takes hard partitions/classifications, this ignores the information about uncertainty which can artificially inflate or deflate the measure of similarity between clustering results. Soft ARI (SARI) is a new ARI style index with a similar interpretation that uses membership probabilities (and also potentially partitions) in its calculation. This talk will discuss how SARI is constructed, its interpretation and comparison to ARI and specific examples of how it can be used to quantify uncertainty in clustering similarity.

1.7 Contributed - Data science: Clusters and networks
Tuesday 5 September – 9am-10am

Network Influence: A New Measure on the Importance of Network Components

Frederick Kin Hing Phoa, Livia Lin-Hsuan Chang
Academia Sinica

A network consists of nodes and edges. For the past 30 years, the study of centrality provided a measure of the importance of nodes, but few provided statistical tests for verification. A recent work on focus centrality proposed a likelihood ratio test to check the validity of node centers, but it was of limited uses with various constraints, including but not limited to, an assumption on undirected networks. This work introduces a new quantity, called Network Influence, to measure of the importance of nodes in either a community or the whole network. The selected network subset can be directed or undirected with self-connection in the structure, while additional information in the form of attributes in nodes and edges are considered in addition. We provide a likelihood ratio test on the network influence to verify if the selected nodes is truly influential to the network. This method is demonstrated in a comprehensive study of a big network data, the Web of Science, revealing many interesting phenomena in paper citations in various research fields. A cross-study between the citation network of statistics and those of various research fields provides a spectrum of the technicality on the use of statistics in each research fields. The result of this cross-study reveals the true importance of statistics, as a fundamental research tool for data analysis, towards all scientific researches, which is difficult to explicitly visualize in other citation indicators like impact factors or traditional centrality.

1.8 Contributed - Medical statistics: **Predictive and prognostic modelling**
Tuesday 5 September – 9am-10am

Predictive modelling for healthcare associated infections: machine learning v.s. classical approaches

Jiafeng Pan, Kim Kavanagh, Chris Robertson, Marion Bennie, Charis Marwick, Colin McCowan
University of Strathclyde

Objectives

The use of patient history data to support individual patient management could be delivered by building robust risk assessment tools onto administrative linked NHS data. We consider creation of prediction models for the risk of acquiring a healthcare associated infection (HAI) at the point of healthcare interaction which could aid clinical decision making.

Approach

We demonstrate this using the HAI Clostridium difficile (CDI). Using linked national individual level data on community prescribing, hospitalisations, infections and death records we extracted all cases of CDI and compared to matched population-based controls to examine the impact of various risk factors including antibiotic exposure, to the risk of CDI acquisition. Predictive models were built using classical conditional logistic regression and machine learning techniques (lasso and random forest) and their ability to predict CDI were assessed using cross-validation.

Results

In the period 2010-2013 there were 1446 CDI cases with 7964 matched controls. Regression modelling highlighted previous cumulative prescribing as a predictive factor (1-7 days exposure OR=3.8 rising to OR=17.9 for 29+ days) and achieved sensitivity 69%, specificity 77% and AUC 79%. Random forest highlighted comorbidities being of dominating importance. Optimal lasso model required more factors than reduced regression model (29 vs. 8). The machine learning techniques improved predictability (lasso/ random forest: sensitivity 71%/70%, specificity 80%/81% and AUC 81%/82%).

Conclusion

Machine learning allows inclusion of highly correlated variables which would cause collinearity in the regression, and it improves predictability although only marginally. The output is hard to interpret clinically therefore is not useful for simple clinical decision support but would be useful if embedded within systems. Contrastingly, regression outputs facilitate the understanding of risk factors highlighting potentially modifiable behaviour. Compared to lasso, using backwards selection includes fewer variables but the manual nature of the algorithm is less adaptive if patterns change over time.

1.8 Contributed - Medical statistics: Predictive and prognostic modelling
Tuesday 5 September – 9am-10am

Estimating risk of seizure recurrence for people with epilepsy

Laura Bonnett, Anthony Marson, Jane Hutton, Catrin Tudur Smith
University of Liverpool

Prognostic models within epilepsy tend to model time to a specific event such as 12-month remission from seizures or first seizure after randomisation via Cox's proportional hazards model. Although such models are simple to fit and easy to interpret clinically, they fail to account for the recurrent nature of seizures which define epilepsy. The Prentice, Williams and Peterson - Counting Process (PWP-CP) model was therefore used to estimate risk of seizure recurrence including all seizures across the whole follow-up period, not just those occurring prior to the specified time point. This is the first time such a model has been applied to epilepsy data. We show that this model demonstrates improved statistical power and enables a better understanding of treatment effects within a clinical trial.

The results from the PWP-CP model were similar to those for conventional Cox models for related fixed time point outcomes. In general, the direction of effect was consistent. However the confidence intervals obtained via the PWP-CP model tended to be narrower due to the increase in statistical power.

This work suggests that the PWP-CP model is appropriate for modelling seizure recurrence by accounting for the clustering of seizures experienced by patients with epilepsy. Further work is required to validate the model and demonstrate its increased statistical power in alternative data. Such a model may be worth considering when designing future clinical trials in medical conditions typified by recurrent events to ensure improve efficiency and statistical power.

1.8 Contributed - Medical statistics: Predictive and prognostic modelling
Tuesday 5 September – 9am-10am

Developing predictive models for severe postoperative complications in cardiac patients

Linda Lapp, Matt-Mouley Bouamrane, Kimberley Kavanagh, Stefan Schraag
University of Strathclyde

Postoperative complications are known to be significantly associated with mortality and morbidity. Severe postoperative complications after cardiac surgery can have a significant impact on patient's quality of life, hospital length of stay and healthcare costs.

All patients in Golden Jubilee National Hospital undergoing cardiac surgery between April 1, 2012 and March 31, 2016 were investigated. The clinical audit database CaTHI consisted of preoperative variables describing patient characteristics, comorbidities, general cardiac status and surgery, and outcomes such as death, hospital length of stay, ICU hours, and presence of complications. A model to predict severe postoperative complications was developed, using logistic regression and performance assessed using receiver operating characteristic (ROC) curves. The predictive ability was compared to the commonly used preoperative risk of mortality model logistic EuroSCORE (LES).

Of 3700 analysed admissions, 59.7% had CABG, 26.4% aortic valve, and 13.9% combined CABG and aortic valve surgery. According to preoperatively calculated LES, 2.5% were high-risk patients (LES >20). The prevalence of severe complications was 5.0% (95% CI 4.3-5.7%). The locally developed prediction model for severe complications consisted of age, gender, diabetes, left ventricular function, previous operations, hypertension history, active endocarditis and previous myocardial infarction. The area under the ROC curve (AUC) of the model was 0.667 with 72.1% sensitivity and 56.6% specificity, with positive (PPV) and negative predictive values (NPV) being 2.5% and 92.0%, respectively. The AUC of LES model predicting severe postoperative complications was 0.672 with 60.7% sensitivity and 66.0% specificity, PPV of 3.0% and NPV of 91.4%.

The high NPVs of both models and high sensitivity of the local model show that both could be used to identify patients without severe complications in order to allocate resources accordingly. A model predicting severe complications could help reducing overall costs of care and improve patients' quality of life.

Plenary 2
Tuesday 5 September – 10.20am-11.20am

Surfaces, shapes and anatomy

Adrian Bowman
University of Glasgow

Methods for imaging the surfaces of three-dimensional objects are now widely available and surface data are routinely collected in many areas of science. This generates interesting questions about how to define and quantify shape in this context, how to estimate the key characteristics, and how to use shape models in the context of scientific studies. This talk will have a strong application focus, referring to data principally from medicine and biology. Examples include the need to quantify the effects of surgery, investigation of issues in human biological development, and the impact of environmental change on the shapes of organisms. The human face is a shape of particular interest. Models for shape change over time will be illustrated, at short scale in facial animation, medium scale in individual growth patterns, and very long scale in phylogenetic studies.

2.1 Invited - Clinical trial estimands - moving from definition to estimation
Tuesday 5 September – 11.50am-1.10pm

Inference about estimands other than ITT in randomised trials: importing methods from causal inference

Rhian Daniel
Cardiff University

In theory, randomisation makes drawing causal inferences from RCTs an almost trivial exercise, with the one caveat (sometimes called the intent-to-treat principle) that the causal estimand being targeted is the effect of *being assigned* to one treatment arm versus the other, since it is assignment to treatment arm that is randomised.

Those involved in the analysis and interpretation of data from RCTs are often interested in estimands *other than* the effect of being assigned to treatment, however. For example, they may be interested in the effect of actually taking the assigned treatment, or in the effect of taking the treatment in a hypothetical world in which some other feature (side effect, initiation of rescue medication, etc) is avoided.

This interest in other estimands has for too long remained implicit; for example, in some longitudinal RCTs, the fact that what was being targeted was not the ITT estimand was only cryptically implied by the manner in which missing data were imputed. In response to this, the Steering Committee of the International Council for Harmonization (ICH) embarked on a detailed consultation on the range of possible estimands for RCTs, and has prepared an addendum to ICH guideline E9, highlighting the need for clarity on the choice of estimand(s) in RCTs, and suitable methods for their subsequent estimation.

In the subfield of statistics known as "causal inference", many useful methods (known as "g-methods") have been developed by Jamie Robins and colleagues to deal with the issue of time-dependent confounding in longitudinal observational studies. In this talk, I will introduce these methods, showing their relevance to the analysis of RCTs when estimands other than the ITT estimand are of interest.

2.1 Invited - Clinical trial estimands - moving from definition to estimation
Tuesday 5 September – 11.50am-1.10pm

Estimands in Clinical Trials – Broadening the Perspective

Mouna Akacha
Novartis Pharma AG

Defining the estimand of interest in a clinical trial is crucial to align its planning, design, conduct, analysis, and interpretation. The need for more precise specifications of estimands was highlighted by the Steering Committee of the International Council for Harmonization (ICH) in 2014, which endorsed a Concept Paper with the goal of developing a new regulatory guidance, suggested to be an addendum to ICH guideline E9.

The estimand discussions have highlighted that some established paradigms in the pharmaceutical industry, at least when it comes to the statistical concepts and principles, may need to be adapted and expanded in line with the scientific questions that are of interest to patients, prescribers, regulators, payers, and the sponsors. We will use two case studies to illustrate the benefits of the estimand framework in facilitating the specification of the scientific question. Moreover, we will touch upon associated statistical methods and the required assumptions.

2.1 Invited - Clinical trial estimands - moving from definition to estimation
Tuesday 5 September – 11.50am-1.10pm

Estimands – What are the key concepts in the ICH E9 Addendum and what challenges and opportunities does this create for new research?

David Wright
AstraZeneca

In trials things happen that complicate the interpretation of the trial's results. For example, a patient discontinues randomisation treatment becuase they cannot tolerate it and then receives an alternative treatment. The intention to treat (ITT) analysis compares outcomes irrespective of the occurrence of such 'intercurrent' events. The ITT analysis answers a question that may not be of primary interest to all stakeholders. Clinicians and patients might for example be interested in the effect that the patient might receive if they are able to tolerate the treatment and stay on it. Such discussions make clear that there are a number of possible scientific estimands of interest in a randomised trial.

Historically in trials it was often not clear what the estimand of interest was. Instead, the statistical analysis would specify how the analysis would handle things such as treatment switching or discontinuation. From the analysis, one might then infer backwards what the estimand is supposed to be. More recently it has been recognised that trials should instead upfront define what the scientific estimand of interest is. From this the statistical analysis should then be chosen such that it targets the chosen estimand, ideally under minimal, or at least plausible assumptions. To aid in this process, the ICH E9 addendum aims to promote harmonised standards on the choice of estimand in clinical trials. It will describe a framework for defining estimands, and also for planning and conducting analyses to assess sensitivity to assumptions.

This talk will explain the framework described in the addendum, mention some benefits from more precisely specifying the primary question of interest and highlight two areas where further research is required,

1. Where estimation of an estimand of interest is currently not possible
2. Where it can be debated what is the most appropriate choice of estimation for a particular estimand.

2.2 Invited - From better data to better decisions to better lives
Tuesday 5 September – 11.50am-1.10pm

From better data to better decisions to better lives

Hugh Stickland
Office for National Statistics

Matthew Powell
Oxford Policy Management

Statistics in themselves have only latent value which is realised when they are actually used to make better decisions that improve welfare. Drawing on examples from the work the UK Office for National Statistics is undertaking to better inform public policy through generating relevant evidence for decision making and on the ways the government of the Punjab is using evidence from new data sources to improve policy and the delivery of health and disaster relief services, this session will showcase some of the challenges and opportunities facing statisticians in analysing and demonstrating the links from better data to better decisions to better lives.

2.3 Invited - Quantifying the changing nature of health inequalities
Tuesday 5 September – 11.50am-1.10pm

Investigation of inequalities in vaccine uptake in Scotland

Chris Robertson, Duncan Lee, Gary Napier, Andrew Lawson, Kevin Pollock, Ross Cameron, Kim Kavanagh
University of Strathclyde

In the United Kingdom the measles, mumps and rubella (MMR) vaccine is offered to all children aged 1, with a second dose before school at age 5. High levels of uptake are crucial to prevent large outbreaks of measles. A substantial decrease in UK vaccination rates was observed in the early years of the 21st century following publication of a, subsequently retracted, article linking the MMR vaccine to autism.

This talk describes a spatio-temporal Bayesian hierarchical model with accompanying software (the R package CARBayesST) to simultaneously address three key epidemiological questions about vaccination rates: (i) what impact did the controversy have on the overall temporal trend in vaccination rates in Scotland; (ii) did the magnitude of the spatial inequality in measles susceptibility in Scotland increase due to the MMR vaccination scare; and (iii) are there any covariate effects, such as deprivation, that impacted on measles susceptibility in Scotland. The model is applied to measles susceptibility data in Scotland among a series of cohorts of children who were aged 2-4, in the years 1998 to 2014.

A second example investigates the uptake of HPV vaccination among girls aged 12-17 in Scotland since 2008 and its impact on cervical disease. This vaccine prevents cervical cancer that has a higher incidence among women living in more socially deprived areas. The HPV vaccine is delivered in schools and factors affecting the uptake of this vaccine are investigated using the same model. This analysis demonstrates low levels of inequality in vaccine uptake and that while the HPV vaccine is associated with significant reductions in both low- and high-grade cervical disease for all deprivation categories, the effect on high-grade disease was most profound among women living in the most deprived communities.

2.3 Invited - Quantifying the changing nature of health inequalities
Tuesday 5 September – 11.50am-1.10pm

A Bayesian Space-Time Model for Clustering Areal Units based on their Disease Trends

Gary Napier, Duncan Lee, Chris Robertson, Andrew Lawson
University of Glasgow

We present a novel general Bayesian hierarchical mixture model for clustering areas based on their temporal trends. Our approach is general in that it allows the user to choose the shape of the temporal trends to include in the model, and examples include linear, general monotonic, and changepoint trends. Inference from the model is based on Metropolis coupled Markov chain Monte Carlo $(MC)^3$ techniques in order to prevent issues pertaining to multimodality often associated with mixture models, with the effectiveness of $(MC)^3$ demonstrated in a simulation study. The model is then applied to hospital admission rates due to respiratory disease in the Greater Glasgow & Clyde Heath Board between 2002 and 2011 to investigate which parts of the city have shown an increased risk, which have shown a decreased risk and which have shown no change. Software for implementing this model will be made freely available as part of the R package CARBayesST.

2.3 Invited - Quantifying the changing nature of health inequalities
Tuesday 5 September – 11.50am-1.10pm

Education and deprivation as explanations of school-level variation in suicidal behaviour in a cohort of 275,420 Scottish school leavers, 2007-2012

Catherine Stewart, Alastair Leyland
MRC/CSO Social & Public Health Sciences Unit, University of Glasgow

Introduction: Health-related outcomes of school pupils are affected by school attended. Such effects may remain after controlling for pupil characteristics, suggesting the school environment (e.g. school policies) affects the health of its pupils. Limited research has also demonstrated a persistence of school effects beyond school-leaving. We investigate school effects in attempted and completed suicide in a population of school leavers in Scotland and whether this is best explained by individual educational attainment or area-based measures of deprivation.

Methods: Education data for school-leavers during 2007-11 were linked with mortality and hospital records from birth until 2012. Educational attainment was measured using tariff points (range=1-120 per subject). Total scores were calculated by summing all points accumulated during school. Area-based measures of deprivation corresponding to school attended and residential address at school-leaving were available. Multilevel discrete-time survival models allowed quantification of variation in suicidal behaviour between date of school-leaving and September 2012 that was attributable to school attended.

Results: In the population of 275,420 leavers, 3155 (1.1%) attempted or completed suicide after school-leaving. In the sex-adjusted model, 2.8% of variation in the outcome was attributable to school. Adjusting for educational attainment accounted for a greater proportion of the total unexplained between-school variation (42%) than deprivation of residential address at school-leaving (15%). Significant between-school variation was still observed on full adjustment for other individual-level educational and health-related factors (σ_u^2=0.064, se=0.014). School-level area deprivation was not significantly associated with the outcome in addition to compositional factors (p=0.246).

Conclusions: A small but significant effect of school on suicidal behaviour persisted beyond school-leaving. Educational attainment was an important predictor of suicidal behaviour and accounted for almost half of the total unexplained variation between schools. As well as targeting poorly performing individuals within schools, consideration should be given to school environment for reducing suicidal behaviour in young adulthood.

2.4 Invited - ScotCen - Is seeing believing? Producing quantitative output for a non-quant audience
Tuesday 5 September – 11.50am-1.10pm

What Scotland Thinks and What UK Thinks – producing bespoke graphs using polling data from the EU referendum

John Curtice
ScotCen Social Research and University of Strathclyde

Referendums and elections produce a welter of data on public attitudes, the significance and interpretation of which can become the subject of claim and counter-claim as the protagonists attempt to show that public opinion is on their side. As a result, it can be difficult – but also important - for the journalist, financial analyst or interested layperson to form in a timely fashion their own judgement about the latest polling information. Building on the experience of providing such a service during the Scottish independence referendum (whatscotlandthinks.org), during the EU referendum NatCen developed and maintained a web site, whatukthinks.org/eu, that provided a comprehensive searchable database of polling data on attitudes of relevance to the referendum debate together with impartial analysis of the latest polling information, both in the form of blogs and longer analysis papers. In so doing it became both a one-stop shop and a gateway to all the key polling data generated during the referendum campaign. Key features of the site include constructing and displaying the time series being generated by the polls, making all data available in both tabular and graphical form (in both cases downloadable), and discussion of the potential substantive implications of the methodological choices bring made by different polls – in each case with a view to making such information as accessible as possible to users who may not have enjoyed much statistical or methodological training. The paper considers the value of such an exercise, the challenges involved in delivering it, and the role that the site played during the EU referendum campaign (and, following the decision to leave the EU, in the subsequent debate about the shape of Brexit!).

2.4 Invited - ScotCen - Is seeing believing? Producing quantitative output for a non-quant audience
Tuesday 5 September – 11.50am-1.10pm

‘Realigning Children’s Services’: Development of a data visualisation tool for Community Planning Partnerships

Stephen Hinchliffe
ScotCen Social Research

The Realigning Children’s Services programme has been working with Community Planning Partnerships (CPPs) in Scotland to support communities to make better decisions using high quality data on local need, to improve the lives of children in their area.

As part of this programme, a number of surveys have been commissioned, with data from these being linked with administrative data. In order to make these data accessible to a range of decision makers within the CPPs, an interactive data visualisation tool is under development. This will allow users with little experience of quantitative data to explore the data for themselves, and to create bespoke maps and charts.

As well as demonstrating the capabilities of such a tool, there will be discussion of the difficulties of using ‘off-the-shelf’ Business Intelligence software (Tableau) with large-scale survey data. These include the technical expertise required to use the software, the difficulties of fitting the type of data we may be familiar with into a package designed for other purposes, the limitations of the software, and the requirements for ensuring the security of the data and the non-disclosive nature of outputs. Because of the large volumes of potential analyses available to users, there is also a trade-off between the freedom offered to users to explore and interpret the data for themselves and the ability to support them in their understanding and use of the outputs.

This presentation will hopefully encourage other analysts to consider whether their own data can be made available to a wider audience in a similar way.

2.5 Invited - Recent developments in high-dimensional data analysis
Tuesday 5 September – 11.50am-1.10pm

A feature distributed framework for large-scale sparse regression

Chenlei Leng, Xiangyu Wang, David Dunson
University of Warwick

Large-scale data with a large number of features are increasingly encountered. This paper presents a framework for sparse high-dimensional linear regression by distributing features to multiple machines. Our method performs similarly to the infeasible oracle estimator in a centralized setting for which all the data are fitted on a single machine. Remarkably, this performance is achieved for elliptically distributed features including Gaussian variables as a special case, for any heavy tailed noises with a finite second moment, for sparse and weakly sparse signals, and for most popular sparse regression methods. Rather surprisingly, we show that a lower bound of the convergence rate of the resulting estimator does NOT depend on the number of machines. Extensive numerical studies are presented to illustrate its competitive performance.

2.5 Invited - Recent developments in high-dimensional data analysis
Tuesday 5 September – 11.50am-1.10pm

Long-Range Dependent Curve Time Series

Degui Li, Peter M. Robinson, Hanlin Shang
University of York

We introduce methods and theory for functional time series with long-range dependence. The temporal sum of the curve process is shown to be asymptotically normally distributed. We show that the conditions for this cover a functional version of fractionally integrated autoregressive moving averages. We also construct an estimate of the long-run covariance function, which we use, via functional principal component analysis, in estimating the orthonormal functions spanning the dominant sub-space of the curves. In a more general, semiparametric context, we propose an estimate of the memory parameter, and derive its consistency result. A Monte-Carlo study of finite-sample performance is included, along with two empirical applications. The first of these finds a degree of stability and persistence in intra-day stock returns. The second finds similarity in the extent of long memory in age-specific fertility rates across some developed countries.

2.5 Invited - Recent developments in high-dimensional data analysis
Tuesday 5 September – 11.50am-1.10pm

Functional Covariance Models on High Dimensional Functional Data

Xinghao Qiao, Cheng Qian, Chenlei Leng
London School of Economics

Some recent efforts have been devoted to modelling dynamic covariance matrices for independent but non-identically distributed scalar data. In this paper, we extend the covariance models concept to describe the covariance relationships among \$p\$ random functions, each of which can be represented using a functional principal components expansion. We propose functional covariance models to describe the covariance dynamics, where principal component scores characterize the global covariance structure and principal component functions further lead to the functional representation of the covariance features. We then apply either entry-wise thresholding/banding to the estimated functional covariance matrix or blockwise thresholding/banding to the sample covariance matrix of estimated principal component scores for achieving uniform consistency results. Our theoretical results demonstrate the non-asymptotic error rates and support recovery properties of our proposed model even in the high dimensional large \$p\$ small \$n\$ scenario. Finally, we illustrate the sample performance of our approaches through a series of simulations and two real world data examples.

2.6 Invited - Statistical literacy: Understanding & communication of statistics
Tuesday 5 September – 11.50am-1.10pm

Statistical literacy approaches in the UK – progress to date

Scott Keir
Royal Statistical Society

The RSS works with its members and wider stakeholders to improve statistical literacy in the UK through a wide range of activities and policy work. Scott will provide a brief overview of these approaches, mapping them to four broad areas, each of which has a different theoretical background and intended purpose: Statistical literacy at home and in everyday life; Statistical literacy in the world of work; Statistics in culture; Public input into the development of statistics and statistical techniques and tools. He will highlight recent developments in formal education at school and university, and discuss activities that have improved, and are improving, statistical literacy in professions including the media and Government.

2.6 Invited - Statistical literacy: Understanding & communication of statistics
Tuesday 5 September – 11.50am-1.10pm

How can we best help people be more statistically literate?

David Spiegelhalter
University of Cambridge

Odds ratios, confidence intervals, significance, effect size. Stats communication is full of terms that are important but baffling to many 'outsiders' who nevertheless are interested in the basic message. How can we help others both understand what is meant, and then translate the ideas for their own audiences? I shall suggest we need a resource that explains statistical concepts at many different levels, from tweet-length to the full technical stuff, and that provides useful phrases that can be adapted to various circumstances. This will need a substantial collaborative effort.

2.6 Invited - Statistical literacy: Understanding & communication of statistics
Tuesday 5 September – 11.50am-1.10pm

Encouraging and Improving Statistical Literacy in the U.S. Media"

Rebecca Goldin
George Mason University

The media plays an essential role in promoting the role and the impact of quantitative reasoning and statistical literacy, as the adult public has little access or interest in formal statistical education. Promoting a statistically literate society requires careful consideration the way that journalists, writers, bloggers, and other media creators themselves understand and discuss statistical content. Conversely, the scientific and statistical discussions need to be leveled to a public audience, which will arguably have more influence than the scientific community on the resulting impact of scientific work. I will discuss some of the efforts STATS has undertaken to facilitate journalists' learning of quantitative tools, to provide access for the media to the statistical community as a resource, and to create more widespread recognition of the value of writing about numbers in meaningful ways by both journalists and scientists. Language and conceptual impediments can be overcome with careful consideration of context and resources.

2.7 Invited - Data Science of urban movement
Tuesday 5 September – 11.50am-1.10pm

Exploring the relationship between Strava cyclists and all cyclists

Mark Livingston, Jinhyun Hong, David Mcarthur, Kirstie English
University of Glasgow

Little is known about cyclists except from very patchy data collected from a few cycling sensors, from censuses or from surveys. Data that can evaluate current cycling infrastructure or appraise new infrastructure investment is not available for most cities in the UK. Cycle counts from the Strava activity tracking app have been shown to be highly correlated over long periods when compared with bike counters on funnel points like bridges but little has been done to compare over shorter periods for many points examining factors like: the Geography of cities; Weather; time of day. These factors might affect the correlations between Strava and counts of all cyclists. This research explores the differences between Strava counts and counts in Glasgow City Council's annual cordon count and the extent that these data can be reliably used to predict all cycling usage.

Cordon counts are carried out on two days every September, and we compare these complete counts of cyclists (2013, 2014 and 2015 cordon counts) to the subset of Strava cyclists (2013, 2014 and 2015 Strava data). Where the numbers of cyclist are high we found strong correlations with the Strava data. At specific points in the cordon where cycling numbers are low correlations are poor requiring aggregation with nearby cordon points. Correlations are high for different time periods and for single or multiple days. Using Negative Binomial modelling we find significant changes in the relationship between Strava cyclists and the cycle counts for: Year; commuting times; and Geography. We also find significant interactions between Strava cyclists and: year; commuting; and geography which any future prediction model would have to account for. Our study suggests that Strava data could be used as a proxy for predicting cycling numbers in cities where these numbers reach certain thresholds and other relationships like increasing cycling numbers are also accounted for.

2.7 Invited - Data Science of urban movement
Tuesday 5 September – 11.50am-1.10pm

Will using the Internet while travelling reduce future car ownership rates of Millennials?

Jinhyun Hong, David Mcarthur
University of Glasgow

New technologies have significant effects on travel behaviour, attitudes, habits and potentially future travel demand. Effects may be more prominent for Millennials. Little empirical research has investigated these relationships, mainly due to data limitations. This study focuses on the potential influence of using the Internet while travelling on Millennials' plans for car ownership. We examine two questions: does using the Internet while travelling influence trip frequencies?; and does it affect Millennials' intention to purchase a car? Results suggest that Internet use while travelling is positively associated with travel demand and the intention to purchase a car in the near future.

2.7 Invited - Data Science of urban movement
Tuesday 5 September – 11.50am-1.10pm

Utilising Strava Metro data for Research of Urban Transport and Health

Yeran Sun
University of Glasgow

With the development of Information and communications technology, crowdsourced geographic information is playing a large role in studies of transport and public health. Recently, a popular online social networking Strava dedicated to tracking athletic activity (cycling and running) offers aggregated cycling activities data at a large spatial scale (street level and intersection level). With fine spatial granularity Strava Metro data tends to promote studies of cycling and health at a large spatial scale. Moreover, Strava Metro data also distinguish type of cycling activity (commuting or recreational). Accordingly, number of commuting cycling activities and number of recreational cycling activities in a street or an intersection are known. In this case, Strava Metro data offers a good opportunity for the research associated with cycling purpose. In this paper, to demonstrate potential of Strava Metro data in research of transport and health, we assess and map cyclists' exposure to air pollution across Glasgow using Strava Metro data. We will discuss advantages, limitations and outlook of Strava data in studies of transportation and traffic-related health effects.

2.7 Invited - Data Science of urban movement
Tuesday 5 September – 11.50am-1.10pm

Assessing the relationship between social capital and active travel

Prachi Bhatnagar
University of Oxford

Objectives

Inequalities in physical activity levels and the health outcomes that physical inactivity leads to are highly prevalent. Social capital is associated with increased physical activity, but little is known about how this differs by demographic group. We aimed to describe how active travel varies by demographics, social capital and perceptions of the built environment, and assess how these are associated with active travel.

Methods

The integrated Multimedia City Data survey (iMCD) contains a linked Glasgow-based survey and travel diaries, which together provide measures of active travel, social capital, demographic variables and perceptions of the neighbourhood environment. The travel diaries were merged with the iMCD survey, resulting in a final data set of 2,052 observations. Chi square tests were used to test for differences in active travel between groups and logistic regression was used to assess the impact of social capital and perceptions of the built environment on the odds of travelling in an active way.

Results

There were significant differences in active travel between age-groups ($p<0.00$), income groups ($p<0.00$) and educational attainment ($p=0.01$) but not between genders ($p=0.28$). For social capital, there were differences in active travel between those who exercise regularly, but also for those regularly attending concerts and using public libraries. Socioeconomic status measures were inversely associated with the odds of active travel. Rating the area highly significantly reduced the odds of travelling in an active way (0.78), but exercising and visiting public libraries increased the odds (1.17 and 1.38 respectively). Adjusting for socioeconomic status measures increased the effect of library use but did not affect other variables.

Conclusions

These findings indicate that, higher levels of social capital indicators may be associated with increased levels of active travel in Glasgow. Future research should investigate how objective measures of neighbourhood and social capital interact to influence active travel.

3.1 Invited - Survival analysis: beyond proportional hazards
Tuesday 5 September – 2.10pm-3.30pm

Adjusting survival time estimates in the presence of treatment switching

Nick Latimer
University of Sheffield

Synthesis of survival data

Suzanne Freeman
University of Leicester

Modelling approaches to the extrapolation of survival curves

Beth Woods
University of York

The session will consider some of the statistical challenges facing analysts when trying to estimate the comparative effectiveness and cost-effectiveness of treatments that may improve survival. These analyses require statistical methods that go beyond the estimation of hazard ratios within trials. They require the extrapolation of survival curves, the adjustment for differences between the trial and clinical populations, and the synthesis of data from multiple studies. These analyses are challenging - as Yogi Berra said “It's tough to make predictions, especially about the future”. However, given the high costs of certain new therapies for cancer and other conditions, they are essential to ensure the efficient allocation of limited healthcare resources and appropriate incentives for technology developers.

Speakers will discuss statistical approaches to (a) the estimation of mean survival and associated measures; (b) the estimation of treatment effects when the treatment sequences observed in trials do not match those expected in clinical practice, for example when trial subjects switch from comparator to experimental treatment upon progression; and (c) the synthesis of data from multiple studies where the comparison of treatment effects may be confounded by differences between trials.

3.2 Invited - Update from the UK Statistics Authority - Code of Practice Review and Post-Brexit round-up
Tuesday 5 September – 2.10pm-3.30pm

Update from the UK Statistics Authority - Code of Practice Review and Post-Brexit round-up

Penny Babb
Office for Statistics Regulation

The Office for Statistics Regulation is consulting on a new edition of the *Code of Practice for Statistics* (‘*Code 2.0*’) between 5 July and 5 October 2017. In this session Penny will outline the ways that we have built on the feedback from our earlier Code Stocktake in designing Code 2.0. He will also explain our ideas for extending the influence of the Code, not just within official statistics but beyond.

There will be an opportunity for Q&A: we would welcome hearing your thoughts about the Code.

https://www.statisticsauthority.gov.uk/osr/code-of-practice/consultation/

3.3 Invited - Compositional data analysis in modern biology and ecology
Tuesday 5 September – 2.10pm-3.30pm

Compositional canonical biplots

Jan Graffelman
Universitat Politecnica de Catalunya

Compositional data occur in many fields of science. Some examples of compositional data sets are mineral compositions in geology, public expenditure composition in economy, or the species composition of an ecological community. Compositional data consist of vectors that contain parts of some whole, and are inherently of multivariate nature.

Log-ratio principal component analysis (PCA), proposed by Aitchison (1983), has become a standard exploratory tool in compositional data analysis, and is often one of the first techniques used to analyze a compositional data set. Biplots can be used to visualize the compositional data (Aitchison and Greenacre, 2002).

In some studies different types of compositions are measured simultaneously. The relationships between the different sets of compositions are then of interest. Canonical correlation analysis (CCO) is a classical multivariate method for the study of relationships between two sets of variables, an X set and a Y set, which are both compositional in this context. Different log-ratio transformations can be considered for transforming the compositions. We apply the centered log-ratio transformation (clr) to X and Y prior to canonical analysis. In CCO, the covariance matrices of X and Y variables are inverted, giving structurally singular covariance matrices if the clr transformation is used. This singularity problem can be efficiently dealt with by using generalized inverses. In this contribution (Graffelman et al, 2017) we adapt canonical correlation analysis to the compositional setting, discuss its properties, and present compositional canonical biplots for visualizing the relationships of interest. Some empirical data sets are used to illustrate the results.

References:

Aitchison, J. (1983) Principal component analysis of compositional data. Biometrika, 70(1):57-65.
Aitchison, J. and Greenacre, M. (2002) Biplots of compositional data. Journal of the Royal Statistical Society, Series C (Applied Statistics), 51(4):375-392.
Graffelman, J., Pawlowsky-Glahn, V., Egozcue. J.J. & Buccianti, A. (2017) Compositional Canonical Correlation Analysis. Under review.

3.3 Invited - Compositional data analysis in modern biology and ecology
Tuesday 5 September – 2.10pm-3.30pm

Compositional data analysis on log-ratio coordinates

Javier Palarea-Albaladejo
Biomathematics & Statistics Scotland

Compositional data comprise vectors of non-negative quantities carrying relative information. They are commonly expressed in percentages, concentrations, or equivalent units, with respect to a total which is not necessarily equal for all the samples collected. These particularities, when ignored, have been shown to cause diverse technical and interpretability issues in data analysis. For example, matrix singularity in linear models and multivariate analysis or spurious pairwise correlations between components. Building on the log-ratio methodology introduced by Professor John Aitchison in the early eighties, a significant amount of progress has been made in the recent years in both the theoretical and practical aspects of compositional data analysis. This talk provides an overview of the basics of the log-ratio approach and discusses its use with some of the types of data generated in modern biological sciences.

3.3 Invited - Compositional data analysis in modern biology and ecology
Tuesday 5 September – 2.10pm-3.30pm

Dissimilarity measures to characterize compositions of microbial communities

Glòria Mateu-Figueras, Pepus Daunis-i-Estadella, Josep Antoni Martín-Fernández, Mireia Lopez-Siles
Universitat de Girona

The composition of microbial communities play important roles in biology, ecology and human health disciplines. The analyses of microbial communities aim either (i) to describe communities quantifying the relative abundance of individual microbial taxa, (ii) to characterize how the microbial community change across space, time or in a response to a treatment, or (iii) to compare two or more communities quantifying and understanding differences between them. Distance and dissimilarities measures like Bray-Curtis or UniFract (unweighted and weighted) are commonly used to compute the differences between microbial communities or to apply non-parametric manova tests, dimensionality reduction techniques or clustering methods.

Recently the compositional nature of the microbial data has been discussed and the methodology based on logratios is recommended for the analysis of such data sets. This opens the door to the Aitchison distance as an adequate distance or to other logratio indices, for example, the F-E index recently applied to discriminate amongst different intestinal disorders. It remains an open question to study how much the use of the Aitchison distance would beneficially impact results compared with these more commonly used metrics.

In this work we analyse several distances and dissimilarity measures that are popular in the biological sciences. In particular we show how the changes in size and shape of compositions of microbial communities affect to these measures. We explore its subcompositional coherence and its scale invariance, the two main principles of compositional data methodology. To illustrate the performance of these measures and to compare it with the Aitchison distance or other measures based on logratios, real and simulated data sets are analysed. In particular we use a real data set that contains mucosa-associated *Faecalibacterium prausnitzii* and *Escherichia coli* from the gut microbiota of a Spanish cohort. Using the simulated data set we explore the relationship among the considered measures.

3.3 Invited - Compositional data analysis in modern biology and ecology
Tuesday 5 September – 2.10pm-3.30pm

Phylofactorization: compositional graph partitioning captures evolutionary structure of ecological and microbiome datasets

Alex Washburne
Duke University

Samples of ecological communities are compositions of organisms connected by the evolutionary tree - the tree of life depicting the origin of species. One major challenge of ecological data analysis is to analyze the associations of species' relative abundances with environmental or experimental meta-data in light of the species' evolutionary relatedness. This problem is made especially relevant by microbial sequence-count datasets that contain counts (relative abundances) and DNA sequences that can be used to infer trees independent of the count data, and a desperate need to simplify these microbial big data in a way that is biologically meaningful.

The isometric log-ratio (ILR) transform is a natural choice for changing-variables in light of a sequential binary partition like the evolutionary tree connecting species. In this talk, I will introduce a new exploratory statistical tool, 'phylofactorization', a graph partitioning algorithm which iteratively constructs an ILR basis whose elements correspond to edges on the tree of life - edges along which traits may have evolved. Phylofactorization is a dimensionality reducing tool and a latent variable model, in which latent variables correspond to putative traits and have clear interpretations and implications for empirical follow-up studies, thereby connecting microbial big-data collection with more detailed in vitro microbial physiological studies. Phylofactorization also illustrates an inherently biological branch of mathematics: changing variables to more naturally represent the peculiar geometry of compositions of evolving parts. The use of phylofactorization to classify microbes in the soil, tongue and gut, and to diagnose & understand the microbial ecology of inflammatory bowel disease will be presented. An R package has been developed and is now available with summary tools to allow future research in this field.

3.4 Invited - Area-based Scottish inequality measures for policy and research
Tuesday 5 September – 2.10pm-3.30pm

Carstairs deprivation scores for small areas in Scotland

Denise Brown
MRC/CSO Social and Public Health Sciences Unit, University of Glasgow

Objectives

Carstairs deprivation scores were originally created in 1981 using four Census variables (male unemployment, no car ownership, overcrowding and low social class). As near as possible the same four variables have been used to update Carstairs scores at each successive Census. Scores, historically available at the Postcode Sector level of geography, were also created for other small areas (Data Zones and Output Areas) at the latest update following the Census in 2011. This presentation will discuss the construction of Carstairs scores and its strengths and weaknesses compared to alternative area-based deprivation measures. Carstairs scores are used to examine area-based inequalities in mortality between 1981 and 2011.

Methods

Age-standardised mortality rates for males and females, aged 0-64, were calculated in the 3-year period around each Scottish census in 1981-2011. Areas (Postcode Sectors) were divided into population weighted quintiles of deprivation based on Carstairs deprivation scores.

Results/Conclusions

Rates of all-cause mortality decreased by 43% for females and by 48% for males, aged 0-64, between 1981 and 2011. For males, aged 0-64, all-cause mortality rates decreased most in the least deprived quintile (by 58%; from 422 per 100,000 population in 1981 to 175 per 100,000 in 2011) with the lowest rate of reduction in the most deprived areas (by 37%; from 743 per 100,000 in 1981 to 472 per 100,000 in 2011). Remarkably, all-cause mortality rates in the most deprived quintile in 2011 (472 per 100,000) were still 12% higher than in the least deprived quintile 30 years earlier (422 per 100,000). Although mortality rates have fallen across Scotland over the last 30 years, the faster decrease in less deprived areas has led to increasing inequalities in premature mortality.

3.4 Invited - Area-based Scottish inequality measures for policy and research
Tuesday 5 September – 2.10pm-3.30pm

Using SIMD for describing and targeting inequalities in local areas

Maike Waldmann
Scottish Government

The Scottish Index of Multiple Deprivation (SIMD) is the government's official tool for defining area-based deprivation in Scotland.

SIMD combines more than 30 indicators, describing seven aspects of deprivation: income, employment, health, education, access to services, crime and housing.

SIMD is an overall ranking of Scotland's neighbourhoods. Anyone can find out whether their neighbourhood is more or less deprived compared to others, using the SIMD map at http://simd.scot.

SIMD is also an immensely rich dataset that includes hyper-local data on a range of topics.

In this talk, we will present how the SIMD tool and dataset are used in practice by central and local government, the third sector, and researchers. We will show:

- How the third sector and community groups use SIMD to describe deprivation locally to attract funding
- different ways in which the Scottish Government uses SIMD to direct policy
- how SIMD combined with other data was used to generate insights into child poverty in the Orkney Islands
- how some SIMD indicators can be used to look at change over time
- how to create your own deprivation index using the R code in openSIMD

3.5 Invited - Theoretical advances in Bayesian Nonparametric
Tuesday 5 September – 2.10pm-3.30pm

Constructing stationary time-dependent completely random measures

Jim Griffin, Ilaria Bianchini, Raffaele Argiento
University of Kent

Completely random measures (CRMs) underpin many Bayesian nonparametric priors such as the Dirichlet process or the Indian buffet process. In this talk, I will consider extending these priors to time series problems where a time-dependent nonparametric prior is needed. We consider using the Pitt-Walker construction of stationary time series (Pitt and Walker, 2005) to the exponential CRM family of nonparametric priors (Broderick et al, 2017). This leads to processes with a simple AR(1) structure. Posterior inference is challenging due to the large number of latent parameters and some computational approaches will be described. The proposed processes can be straightforwardly employed to extend CRM-based Bayesian nonparametric models such as feature allocation models to time-dependent data. These process can be applied to problems from modern real life applications in a range of fields from computer science to biology. In particular, we develop a dependent latent feature model for the identification of features in images and a dynamic Poisson factor analysis for topic modelling which are fitted to synthetic and real data.

References:

Broderick, T, Wilson, A. C., and Jordan, M.I. (2017). Posteriors, conjugacy, and exponential families for completely random measures. Bernoulli, to appear.
Pitt, M. K., & Walker, S. G. (2005). Constructing stationary time series models using auxiliary variables with applications. Journal of the American Statistical Association, 100, 554-564.

3.5 Invited - Theoretical advances in Bayesian Nonparametric
Tuesday 5 September – 2.10pm-3.30pm

Dependent nonparametric priors with hierarchical structures

Igor Pruenster
Università Bocconi

Within the framework of partial exchangeability, we consider nonparametric priors based on hierarchical constructions of dependent completely random measures (or suitable transformations thereof). We derive some of their distributional properties including the partition structure and a posterior characterization. These highlight some key features and inferential implications of this class of dependent nonparametric priors and of its popular special case, the hierarchical Dirichlet process. Illustrations deal with prediction within species sampling problems and inference on survival data.

3.5 Invited - Theoretical advances in Bayesian Nonparametric
Tuesday 5 September – 2.10pm-3.30pm

Non-exchangeable random partition model for microclustering

Francois Caron
Oxford

Clustering aims at finding a partition of the data. In a Bayesian framework, this task is addressed by specifying a prior distribution on the partition of the data. Popular models, such as the Chinese Restaurant Process and its two-parameters generalization, rely on some exchangeability assumption; while this assumption may be reasonable for some applications, it has strong implications on the asymptotic properties of the cluster sizes. In fact, as proved in Kingman (1978) and stressed by Miller et al. (2015), exchangeable random partitions imply the linear growth of the cluster sizes, which is not suitable for several applications. We will present a flexible non-exchangeable random partition model, based on completely random measures, which is able to generate partitions whose growth of the clusters sizes is almost surely sublinear. Along with this result, we provide the asymptotic behaviour of the number of clusters and of the proportion of clusters of a given size. Sequential Monte Carlo algorithms are derived for inference and we provide an illustration of the fit of the model on a movie review dataset.

This is joint work with Giuseppe di Benedetto and Yee Whye Teh.

3.5 Invited - Theoretical advances in Bayesian Nonparametric
Tuesday 5 September – 2.10pm-3.30pm

Dependent nonparametric priors with nested structures

Antonio Lijoi
Università Bocconi

The combination of discrete random structures has become an important tool in Bayesian nonparametric modelling. In this talk it is used for the construction of a class of dependent nonparametric priors that are suitably designed to model non exchangeable data. They are obtained by nesting normalized random measures with independent increments. A few relevant distributional properties will be displayed. In particular, it will be seen that with multi-sample data the proposed models are able to capture various forms of dependence ranging from exchangeability to independence across samples. Illustrative examples will provide further insight on these theoretical findings.

3.6 Invited - Promoting statistical literacy in Scottish schools
Tuesday 5 September – 2.10pm-3.30pm

Practice: how statistics is approached within 'numeracy across learning' in schools

Tom Macintyre
University of Edinburgh

A central tenet of the Curriculum for Excellence framework in Scottish education has been the emphasis on 'across learning' opportunities in literacy, numeracy and health and well-being. All practitioners therefore have a responsibility to promote numeracy across learning, with two of the eight numeracy topics being relate to statistical literacy. The curriculum topics on 'Data and Analysis' and 'Chance and Uncertainty' run through the whole school curriculum from Early Level through to Senior Phase. Some learners will choose to study explicit courses in Statistics, such as the one outlined by Young and Reilly in this session, but others will be reliant on developing their competence and confidence in statistical skills through interdisciplinary studies.

Questions that will be discussed in this session include: How well have learners mastered those aspects of numeracy in recent years? How confident are practitioners in supporting and delivering topics related to statistical literacy? What skills and practices need to be developed for young people to be statistically literate in the 21st century? In what ways can learners be supported in becoming data literate to participate as active citizens in their local and national communities?

Evidence from the Scottish Survey of Literacy and Numeracy (SSLN, 2015) and data collection through National Improvement Framework (NIF, 2016) will highlight the case for action on improving statistical education in schools and beyond, paving the way for discussion of resources and developments illustrated by other presenters in this Communicating Statistics session on statistical literacy in Scottish schools.

3.6 Invited - Promoting statistical literacy in Scottish schools
Tuesday 5 September – 2.10pm-3.30pm

Resources: Using data in teaching - developing digital skills by incorporating real data in teaching

Cecilia Macintyre, Tom Macintyre
Scottish Government

Census taking is covered in the social subjects curriculum primarily as a way of illustrating demographic trends over time, and contrasts between developing and developed parts of the world. Curriculum for Excellence encourages developing skills in using and manipulating data, and the Census team in Scotland were invited to present to a group of teachers at a workshop on using real data. This event sparked interest and following this a series of workshops were organised to introduce teachers to the wealth of data available on Scotland's Census data explorer. http://www.scotlandscensus.gov.uk/

This talk will outline the approach which was taken to introduce the teachers to the possibilties in the data, outline the links to the Curriculum for Excellence, and describe the resources which are currently available for teachers to use.

Future plans for disseminating the resources, and evaluating its impact will be outlined. Also the challenges in encouraging update of this resource in a crowded curriculum will be discussed.

3.6 Invited - Promoting statistical literacy in Scottish schools
Tuesday 5 September – 2.10pm-3.30pm

Scottish Qualifications Authority Statistics Award

John Reilly, David Young
University of Strathclyde

In a world where more and more data is being collected, there is a need for young people to learn how to analyse and report data, and to apply these skills to real world problems. Statistics was generally taught in the lower school as part of the mathematics curriculum. There is an Advanced Higher in Statistics but this does not have a high uptake and tends only to be taught in schools where a particular staff member has a personal interest in the subject. The course content is more traditional, with the emphasis on understanding the concepts, rather than on the application to real data. There are clear advantages for pupils who have a sound understanding of data, and the ability to use statistics to make sense of it, both in the workplace and across various disciplines in Higher Education.

The Statistics Award (SCQF level 6) was developed in collaboration with the SQA and Strathclyde University in an attempt to provide an opportunity for pupils to learn these skills. The emphasis of the course is on the application of statistics to real-life data and interpretation of the results in the context of the scientific study. Initially the course was designed to use Minitab for the practical sessions, however this proved problematic for schools on a limited budget. Currently the course materials are being converted to R as the preferred software package.

As a resource for candidates, a text book is in the final stages of preparation. An online resource for teachers is also being developed which includes lecture slides, examples, data and solutions with videos illustrating how the analyses should be conducted using both R and Minitab.

3.7 Invited - Grand challenges in Data-Centric Engineering
Tuesday 5 September – 2.10pm-3.30pm

Statistical Methods for Instrumented Infrastructure

Din-Houn Lau
Imperial College London

The statistical analysis and methodologies applied to data acquired from instrumented infrastructures is a key technology that Data Science should be currently developing. The use of statistical analysis and methods will improve productivity of networks through better understanding of their current infrastructure.

As an example of the current stage of uptake of this technology: aeroplanes are already instrumented to maintain their productivity i.e. keep engine running for as long as possible. Further, personal and local monitoring of small systems is already being used today; for instance, electricity smart meters. Energy companies provide customers with smart meters – an instrument to measure the electricity consumption of a household. This device allows for better management and control of the system. This is a real-world example of instrumentation. We intend to extend the idea of instrumentation to larger networks such as water networks and transport.

In this presentation, we present a new exciting project: Smart Instrumented Bridges – this project is part of the Alan Turing Institute programme on Data-Centric engineering funded by the Lloyd’s register foundation. This project involves monitoring instrumented rail bridges, and developing statistical tools to accurately assess their *structural health* over time. These statistical methods could potentially avoid heavy-handed and costly interventions which would be of interest to structural engineers and rail companies who maintain such bridges.

In the presentation, we describe the engineering and statistical questions that we can address using data obtained from instrumented structures, describe features of the data and use statistical methods that provide some insight into the solutions.

3.7 Invited - Grand challenges in Data-Centric Engineering
Tuesday 5 September – 2.10pm-3.30pm

Prediction and Tomography of the London Underground

Ricardo Silva, Nicolo Colombo, Soong M Kang, Edoardo Airoldi
UCL / The Alan Turing Institute

This talk addresses questions of prediction and latent structure inference in modern urban train networks from anonymised passenger behaviour. The purpose is to better understand different patterns of usage so that transport authorities can better serve their users and be more readily prepared to respond to unexpected disruptions. All models presented are scalable to millions of origin-destination records and validated with anonymised statistics of the London Underground. In the first part of the talk, we show how predictive models of passenger reaction to unplanned disruptions can generalize from historical occurrences to new situations. This also provides an assessment of station robustness to disruptions in their respective neighbourhoods. The second part of the presentation addresses the classical network tomography problem of inferring local traffic given origin-destination observations. We build a scalable probabilistic model that exploits input-output information to estimate the unobserved link/station loads and the user's path preferences. Based on the reconstruction of the users' travel time distribution, the model is flexible enough to capture possible different path-choice strategies and the variability among users travelling on similar paths at similar times.

3.7 Invited - Grand challenges in Data-Centric Engineering
Tuesday 5 September – 2.10pm-3.30pm

Predictive monitoring of Gas Turbines

Catalina Vallejos, Anthony Latimer, Jason McGinty
The Alan Turing Institute

Failures within complex engineering systems can have high-consequences in terms of efficiency and safety. Unlike reactive strategies — where maintenance is scheduled after a failure is observed — *predictive monitoring* aims to continuously assess the health of the system, understanding and anticipating failures. This aims to deliver a more efficient maintenance schedule by improving the resilience of these complex structures. This project focuses on data-driven approaches for predictive monitoring of gas turbine engines, combining historic failure data together with *condition monitoring data* — i.e. the information that is continuously recorded by a limited number of sensors that are located inside the engine. We describe some of the challenges that arise within this context and explore the use of *reliability models* as a tool to dynamically quantify the risk of the system and to predict time-to-failure.

3.8 Invited - The role of Sports Science in recent Olympic and Paralympic success
Tuesday 5 September – 2.10pm-3.30pm

Tracking Medal Success at the Rio 2016 Paralympic Games

Oliver Summers
UK Sport

The Paralympic Games had around 1600 medals on offer, with 264 British Athletes competing for 349 medals in many different sports and disciplines. With the media's attention on Great Britain and future funding at stake, the ability to answer the following questions was critical:

- How do we know we are on track for 121 medals, the target set by UK Sport?
- When, where and who are the next medal shots?
- Was today a good or bad day?

Therefore in the lead up to the Games, ParalympicsGB collected medal probabilities for every single entry and ran Monte Carlo simulations to assess the likelihood of total medal hauls. For in-Games analysis, the medal probabilities were mapped to the event schedule enabling the calculation of required daily medals on the road to 121. By monitoring how the situation changed on a daily basis, senior leaders from UK Sport, the British Paralympic Association and the English Institute of Sport were well informed on the state of the British campaign, where the next medals were expected and, importantly, aligned on the messaging going out to the press.

With 64 gold and 147 total medals, surpassing the target of 121 medals set by UK Sport, the 2016 Paralympic Games was an incredible success for ParalympicsGB.

Plenary 3 - Campion (President's Invited) Lecture -
Tuesday 5 September – 3.40pm-4.40pm

The Media's Love-Hate Relationship with Statistics: Challenges in Communication

Rebecca Goldin
George Mason University

News accounts are filled with numbers and implicit advice. They also tell a story. Are kids doing more drugs than they used to? Are polls bad at predicting elections? Are boys better at mathematics than girls? Does tax policy influence drug abuse? Can brain scans read our minds? In contexts as diverse as criminal courts, opinion surveys, and our personal health, statistics are playing an increasingly vital role, and highlighting the public's need for clear rendering of numerical information.

Yet many media accounts using statistics are misleading or confusing, for reasons as varied as the news accounts themselves. I will share both humorous and serious stories about statistical bloopers in the media, along with insights about the challenges journalists face, both logistical and conceptual. Statistical reasoning can be powerful when we move past politics and morality to clarify what science actually tells us, what it does not, and what it cannot.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Statistical learning approaches for global classification of lake water quality

George Gerogiannis, Ruth O'Donnell, Claire Miller, Marian Scott
School of Mathematics and Statistics, University of Glasgow

In the age of ‘big data’ the quantity of data produced by sensor technologies is increasing at an unprecedented rate. Data structures produced by these new technologies are often complex and in many situations it is of interest to use statistical learning methods in order to explore the underlying structure within such data sets. In a broad spectrum of fields such as finance, business and environmental sciences there is an aim to partition a data set, establishing groups of individuals or objects which have similar features, or which conform to some pre-specified set of classes.

We focus on data which has arisen from the satellite observations of 732 lakes across the globe, specifically lake surface water temperature. While previous studies have often focused on individual lakes, the availability of data derived from earth observation satellites has enabled us to explore lakes on an unprecedented scale and presents a rich area for statistical research. Lakes are considered as sensitive indicators of environmental change which are impacted by both natural and anthropogenic drivers. The potential impact of climate change on freshwater resources are critical, and improved understanding of the observed changes is key to ensure better management of aquatic resources.

This poster will focus on classification of the lake surface water temperature data, comparing a range of methods from traditional techniques such as discriminant analysis to more modern machine learning techniques such as support vector machines with the aim of identifying lake patterns that are temporally coherent.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Latent financial networks projection via correlation tensor decomposition

Giuseppe Brandi
LUISS University

Nowadays, networks represent a widely analysed topic in financial research. Networks' topology is explored and measures of connectivness and centrality are extracted for each stock (node). The estimated network relies on the sample used to build it. Hence, by construction, it is conditioned on time horizon used and the sampling scheme will affect the results at any level of the network analysis, possibly leading to different conclusions. In this paper we propose a method which relies on tensor decomposition allowing to separate the structural part of the financial network, cleaned by time specific events, from its dynamic one. The first object can be used to assess systemic risk while the second one can be used in evaluating idiosyncratic (dynamic) risk. We show that the static representation does not vary (in terms of topology's statistics) across non overlapping samples while the standard networks do.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Mental Health: Data changing lives

Rachel McIlroy CStat
NHS England

For far too long, people of all ages with mental health problems have been stigmatised and marginalised, often experiencing an NHS that treats their minds and bodies separately. Mental health services are chronically underfunded and many people have received no help at all, leading to hundreds of thousands of lives put on hold or ruined, and thousands of tragic and unnecessary deaths.

In the Spring of 2016, the Five Year Forward View for Mental Health and the Prime Minister's 'Challenge on Dementia 2020' were launched, significantly increasing the focus on the provision and delivery of Mental Health services in England. This aims to deliver "parity of esteem", the principle by which mental health must be given equal priority to physical health as enshrined in law by the Health and Social Care Act 2012.

Consistent and reliable data in mental health lags behind other areas of health. There is good information available, but it is not co-ordinated or analysed usefully. This poster will cover the work of NHS England in the following:

- Promoting availability of Mental Health data
- Developing user-friendly tools and dashboards to enable colleagues to have confidence in and really understand their data and the application to decision making
- Identifying other data sources and providing ways of meaningfully connecting these
- Ongoing stakeholder engagement to establish future data needs

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Some considerations for the design of discrete choice experiments raised by a study on peatland restoration

Jacqueline Potts, Klaus Glenk
Biomathematics and Statistics Scotland

A discrete choice experiment has been carried out to assess the public's willingness-to-pay for a programme of peatland restoration in Scotland. Peatlands store carbon, so restoration reduces greenhouse gas emissions and is also beneficial for water quality and wildlife. Each choice set included a baseline option in which no restoration is undertaken, together with two options involving some degree of restoration at a specified cost in the form of increased taxation. The inclusion of a "status quo" option in each choice set is generally recommended in the literature, but has implications for the efficiency of the experimental design. An orthogonal design was used for a pilot study and prior information from this was used to generate a Bayesian efficient design for the main survey. The incorporation of prior information tends to eliminate choice sets with obviously dominant alternatives.

There was evidence that some respondents may have applied heuristic rules such as always choosing the cheapest non-baseline options and ignoring the levels of other attributes. By means of a simulation study, we investigate the effect of a proportion of respondents applying such rules on estimates of willingness-to-pay obtained from models that do not incorporate this non-attendance to attributes, and consider the implications for the design of future choice experiments.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Bayesian Spatial Monotonic Regression for Modelling Weather Related Insurance Claims

Christian Rhorbeck, Deborah Costain, Frigessi Arnoldo
Lancaster University

A flexible modelling framework which allows for spatially varying covariate effects whilst borrowing any ‘local’ information remains an open challenge. Many techniques assume linearity or smoothness in the functional form; restrictions which may not be appropriate due to existing threshold and interaction effects. We aim to model the association between a response and covariates for a set of contiguous regions under the ‘weaker’ assumption of monotonicity; a conjecture which is valid in many applications. This presentation introduces the proposed Bayesian Spatial Monotonic Multiple Regression methodology. The rationale is to estimate a monotonic function of the covariates for each region, and provide a robust approach in the sense that any threshold, and, or interaction effects can be detected, via a random marked point process.

Monotonic regression features in several research areas including functional data analysis and Bayesian inference. However, little research refers to its application to lattice data. The main idea is to smooth the functions spatially to incorporate potential similarities of neighbouring regions. In our method, this is achieved by the definition of a joint prior on the monotonic functions. Each function is represented by a marked point process which allows for high flexibility. Estimates are obtained by reversible jump MCMC and cross-validation. Based on the samples drawn by the algorithm, predictions are obtained by Monte Carlo integration and redundant covariates are detected.

The method is assessed using simulated data and also applied to Norwegian insurance claim data with a view to modelling weather related claim dynamics over the region. Results illustrate that the predictive capacity is good and also the capability for extrapolation and recovering non-linear effects

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Development and validation of a prediction rule for psychiatric hospital readmissions of patients with a diagnosis of psychosis

Maria Vazquez, Hugo Maruri-Aguilar, Ksenija Yeeles, Stephane Morandi, Jorun Rugkasa, Tom Burns
University of Oxford

Multiple studies of psychosis have demonstrated an association between a variety of clinical and demographic patient characteristics and psychiatric hospital readmission. Our aim is to investigate whether these potential predictors retain their predictive value when put together to create a prediction tool for hospital readmission that clinicians could use to guide best care practice for patients with psychosis when discharged from hospital.

A prediction model for psychiatric hospital readmission will be developed using data from an existing single-outcome, parallel arm, non-blinded randomised trial (OCTET). The trial recruited patients aged between 18-65 years, diagnosed with psychosis, registered in one of 60 NHS Trusts providing mental health services in England. Recruitment took place from 10-November-2008 to 22-February-2011. A total of 119/336 (35%) patients were readmitted to hospital in the 12-month follow-up after discharge at baseline (60/167 patients were re-admitted in the control group only). A set of 17 potential predictors has been identified from those found in the literature. Univariate and multivariate logistic regression models will be fitted using a smooth supersaturated polynomial technique which allows for correlations between the predictors. No variable selection will be done and multiple imputation will be used to account for missing values. The final multivariate prediction model will be internally and externally validated, using bootstrap and applying the model to an independent dataset to evaluate its performance, respectively.

We will present the final model and describe any challenges we encounter to obtain it, both clinical and methodological. We will report the independent and adjusted prediction value for each of the pre-selected predictors; the multivariate model's discrimination and calibration, the latter adjusted for optimism; and the model's sensitivity, specificity and area under the receiving operating curve. Findings from the external validation analysis will also be presented.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

R in Hour of Code: a flash course to make statistical programming accessible to anyone, anywhere

Morgan B. Yarker, Michel d. S. Mesquita, Carla A. Vivacqua, André L. S. Pinho, Vidyunmala Veldore
Yarker Consulting, Iowa, USA

This poster will highlight the results of collaboration for a 'flash course' about the basics of R.

The course, ‘An Introduction to R’, was created to support the worldwide Hour of Code, which is an effort to encourage everyone to spend at least one hour each year learning how to code something new in a simplified way. Our free one-hour online tutorial utilizes robust learning theories and educational strategies to encourage participants to learn how to code in R, while simultaneously learning and practicing statistics. Inquiry-based practices were used that encourage students to ask questions, design experiments, evaluate their findings, and reflect on their understanding. The course was delivered through Moodle on our website at m2lab.org

While the duration of the course was short, it provided participants with basic understanding of R, how to perform basic statistical and plotting functions in R, and the opportunity to analyze real-world data throughout the course. The examples and data provided were from climate science. The final tutorial also provided a problem-solving mini-project as well as a list of additional resources for further learning, such as reading netCDF/geodata format.
Participants were asked to assess their own learning through a series of short-answer questions, as well as during an end-of-course survey. For this poster, participant responses are analyzed to assess their perceived successes and struggles as they completed the course. Of the 16 participants who started the course during the Hour of Code week-long event, 11 completed and obtained certificates. Note that we kept the course open to others and up to now, 117 have taken it, but the analysis focuses on the first cohort. Multivariate statistical analysis is performed and qualitative analysis is also discussed. The focus is on the perspective of responses, rather than respondents.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Inference on the Duffing system with the Unscented Kalman Filter and optimization of sigma points

Michela Eugenia Pasetto, Umberto Noè, Alessandra Luati, Dirk Husmeier
University of Bologna

We analyse the deterministic Duffing process, which describes a periodically forced oscillator, and a characteristic feature is its chaotic behaviour. To infer the parameters of the process, we apply the Unscented Kalman Filter (UKF) algorithm. The UKF relies on the so-called sigma points to obtain the predictive and filtering distributions. The sigma points location is parametrised by three scalar values, and these parameters are heuristically set by the algorithm. The positioning of the sigma points affects the overall inference performance of the UKF and its convergence. In order to improve the convergence of UKF to the true differential equation parameters even in the case of a bad initialization, we optimize the sigma points location using Bayesian optimisation. We also compare the optimization results to the ones obtained by a simpler optimization algorithm on a discretized grid.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Model assessment in cumulative logit ordinal response models: a simulation study

Altea Lorenzo-Arribas, Antony Overstall, Mark Brewer
Biomathematics and Statistics Scotland

Practical issues associated with the discrete nature of the data make model assessment in cumulative logit ordinal response models difficult to implement. Tailored versions of traditional goodness-of-fit measures available for linear models and diagnostic plots of individual residuals are generally unavailable for these models and residual diagnostics are in particular generally accepted as underdeveloped. We assess the behaviour of two types of residuals proposed for ordinal response models (randomised quantile and probability scale residuals) in a simulation study that aims to determine their accuracy in determining lack of fit.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Bayes linear Bayes models and Bayes linear kinematics in medical diagnosis and prognosis

Wael Al-Taie
School of Mathematics and Statistics, Newcastle University

In medical diagnosis or prognosis, we might use information from a number of covariates to make inferences about the underlying condition, predictions about survival or simply a prognostic index. The covariates may be of different types, such as binary, ordinal, continuous, interval censored and so on. The covariates and the variable of interest may be related in various ways. We may wish to be able to make inferences when only a subset of the covariates is observed so relationships between covariates must be modelled. In the Bayesian framework, such a model might suggest the use of Markov Chain Monte Carlo (MCMC) methods but this may be impractical in routine use. We propose an alternative method, using Bayes linear kinematics within a Bayes linear Bayes model in which relationships between the variables are specified through a Bayes linear structure rather than a fully specified joint probability distribution. This is much less computationally demanding, easily allows the use of subsets of covariates and does not require convergence of a MCMC sampler. In earlier work on Bayes linear Bayes models, a conjugate marginal prior has been associated with each covariate. We relax this requirement and allow non-conjugate marginal priors by using one-dimensional numerical integrations. We compare this approach with one using conjugate priors and with a Bayesian analysis using MCMC and a fully specified joint prior distribution. We illustrate our methods with an application to prognosis for patients with non-Hodgkin's lymphoma.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Bayesian Piecewise Constant Hazard Models using Integrated Nested Laplace Approximation (INLA)

Muhammad Irfan bin Abdul Jalal
School of Mathematics and Statistics, Newcastle University

Bayesian survival analysis has benefitted from the introduction of Markov Chain Monte Carlo (MCMC) since the 1990s. However, MCMC has high computational cost and requires tuning and convergence checking. These hamper its usefulness. Integrated Nested Laplace Approximation (INLA) is a practicable alternative to MCMC due to its fast efficient algorithm and straightforward execution to obtain the posterior distributions of relevant survival parameters such as the regression coefficients and Weibull shape parameter. This has been demonstrated in parametric and semi-parametric, piecewise constant, proportional hazard models. Piecewise constant hazard models allow a non-parametric form for the baseline hazard and also allow the coefficients of covariates to change over time. We investigate the application of INLA to piecewise constant hazard models and the extension to allow time-varying covariate effects. We use both hierarchical and autoregressive priors for the baseline log-hazard and covariate effects and compare the results with those obtained by MCMC. We apply the methods to two data sets, one on patients with non-Hodgkin lymphoma (Scottish and Newcastle Lymphoma Group, SNLG, n = 1391) and one on lung cancer patients (HUSM, Malaysia, n = 313). Priors are based on information from previous studies. We fit piecewise constant hazard models and Weibull proportional hazard models to both data sets using both INLA and MCMC. INLA and MCMC produce almost identical posterior summaries of survival model parameters but the computational time is much less for INLA than for MCMC. We conclude that INLA may serve as a fast alternative to MCMC in computing posterior distributions in piecewise constant hazard models.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

A targeted sampling review of the Foreign Direct Investment surveys

Gemma Clayton, Megan Pope, Jonathan Digby-North
Office for National Statistics

The Office for National Statistics (ONS) collects Foreign Direct Investment (FDI) data via both an Inward (foreign parent companies' investment in UK 'affiliates') and Outward (UK parent companies' investment in foreign 'affiliates') survey, each of which has an Annual and Quarterly version. The target population for the FDI surveys compromise UK based businesses where the investing business (foreign or domestic) receives more than 10% of the voting power.

Two sampling frames are used for FDI – the Worldbase (WB) and Non-Worldbase (NWB). The WB frame is an extract from an external database holding information on which companies own others. The NWB is the main sampling frame and contains the larger businesses in terms of their 'net book value' (total investment position).

The principal estimates ONS publish are earnings, flows and positions at various levels of aggregation. The focus on these data will become increasingly important due to 'Brexit', as we will want to understand if and how investments in or by UK companies are changing and any effects this may have.

Users of FDI statistics include HMRC, the Bank of England, the United Nations and many others. FDI data appear in the UK National Accounts Pink Book (Balance of Payments) which measures a net flow of transactions between the UK and the rest of the world.

As a result of the 2016 National Statistics Quarterly Review of FDI, a number of recommendations were made to review various aspects of the sample design and sampling process. Here, we present the results of investigations conducted into some of the high priority areas, such as the stratification of the surveys and the need for a revised allocation of the sample due to a significant increase in the target population of the surveys.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Waring regression model for count data

Antonio Conde-Sánchez, Ana María Martínez-Rodríguez
Universidad de Jaen

GWRM is a package for fitting and computing Generalized Waring Regression Models that is available at CRAN. It includes functions for fitting the model to count data where the response variable follows a univariate generalized Waring distribution (UGWD) with parameters a, k and rho. These models extend the negative binomial models and they are not part of the generalized linear models. In a GWRM model, the variance may be split into three terms. The first component of this decomposition represents the variability due to randomness and it comes from the underlying Poisson model. The other two components refer to the variability that is not due to randomness but is explained by the presence of liability and proneness, respectively. One of the main drawbacks of using the UGWD is that the parameters a and k are interchangeable, so the estimates are less accurate and the components of variance due to the liability and proneness are indistinguishable. In this work illustrative examples in which k is equal to one (that is, the classical Waring distribution is considered) are shown so the estimates are more precise and their interpretation is easier.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

A new aspect of geometrical data analysis using curvature of the data space and the empirical graph

Kei Kobayashi, Henry Wynn
Keio University, Japan

When data points are distributed (exactly or approximately) on a geodesic metric space such as a Riemannian manifold or polyhedral complex, the curvature plays an important role. A simplest example is uniqueness of the intrinsic means of the sample and empirical distributions which is controlled by the CAT(0) property of the geodesic metric space. Firstly, the author's recent study with Henry Wynn on tuning the geodesic metric of each data space via the CAT(k) property is summarized. We explain a method to transform the original metric in two steps, derived from the intrinsic and extrinsic viewpoints, respectively. In order to simplify the computation, the geodesic metric is discretized and approximated via the empirical graphs. The proposed method is verified by several theorems and numerical experiments. While the CAT(k) property is a kind of curvature that can control uniqueness of the intrinsic means, there are many other kinds of curvatures used in geometry such as Ricci curvature. We evaluate how Ricci curvature of the empirical geodesic subgraph is changed by each transformation of the metric by our method and compare it with the results obtained via the CAT(k) properties.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Avoidable Mortality in England and Wales, 2015

Anne Campbell
Office for National Statistics

Prize winner 2016 Young Statisticians Meeting

Background

Avoidable mortality is an indicator that measures the contribution of healthcare to improvements in population health. It is based on the concept that premature deaths from certain conditions should not occur in the presence of timely and effective healthcare. We present avoidable mortality in England and Wales, 2015.

Methods

We calculated the number of deaths from potentially avoidable causes as a proportion of deaths from all causes to gauge the contribution of avoidable mortality to all-cause mortality. Age-standardised rates (ASRs) were used to assess geographical and cause-specific differences in avoidable mortality. Age-standardised Potential Years of Life Lost (PYLLs) were used to measure the years of life that could have been saved had deaths from avoidable causes not occurred. Avoidable mortality in children and young people was also calculated using a separate avoidable mortality definition.

Results

Nearly a quarter of all deaths in England and Wales in 2015 were from causes considered avoidable; for children and young people this rises to nearly a third of deaths.

Avoidable mortality ASRs were higher in males than in females and were significantly higher in Wales compared to England. Within England there was a north-south divide, with regions in the north of England having higher avoidable mortality rates than those in the south.

Across all age groups, the leading causes of avoidable deaths were from chronic conditions: for females lung cancer was the leading cause whereas for males it was ischaemic heart disease. In children and young people non-chronic conditions such as accidents were most common.

Conclusions

Despite advancements in medical technology and public health policy, a substantial number of deaths still occurred from causes considered avoidable through good quality healthcare and wider public health interventions. Avoidable mortality statistics provide heath planners with an early ‘warning sign’ of potential weaknesses in the healthcare system.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Applying a coherent framework to drug utilisation studies: the use of direct oral anticoagulants in patients with atrial fibrillation in Scotland

Tanja Mueller, Samantha Alvarez-Madrazo, Chris Robertson, Marion Bennie
University of Strathclyde

Background: Information regarding adherence to treatment with direct oral anticoagulants (DOACs) is still limited. Drug utilisation research is commonly being conducted to analyse the usage of drugs in clinical practice, but drug utilisation studies make use of a variety of conceptual definitions and a diverse set of measurements. The aim of this study was therefore two-fold: to report on DOAC use in Scotland; and to advocate the standardisation of drug utilisation methods.

Methods: Retrospective cohort study using linked routinely collected administrative data. Patients include those with a diagnosis of atrial fibrillation (AF) who received a first prescription for a DOAC (dabigatran, rivaroxaban, apixaban) from September 2011 to December 2015. In order to give a valid representation of patients` drug taking behaviour, this study comprises various measures of both discontinuation/persistence and adherence.

Results: 14,811 patients (mean CHA_2DS_2-VASc score 2.93 [SD 1.71], 87.2% with ≥ 5 concomitant medicines) were treated with DOACs for a median of 346 days (IQR 167 – 597). 41.4% discontinued treatment during the study period; however, 57.7% re-initiated DOACs, and persistence after 12 months was 80.6%. Differences between DOACs were observed: discontinuation rates ranged from 34.0% (apixaban) to 75.9% (dabigatran), and 12 months persistence from 61.8% (dabigatran) to 83.6% (apixaban). Adherence to treatment with all DOACs was good: overall DOAC median medication refill adherence (MRA) was 102.3% (IQR 90.1% - 112.5%), and 81.9% of patients had an MRA > 80%.

Conclusions: In Scotland, adherence to DOAC treatment was good. However, discontinuation and persistence rates were variable – although treatment interruptions were often temporary.

To decrease the inconsistencies in drug utilisation methods and facilitate meaningful study comparison, the use of a coherent framework – combining discontinuation, persistence and adherence – and the standardisation of measurements is advocated.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Nowcasting GDP: Comparison of methods to complement T+30 approach in official statistics

Dan A. Rieser
European Commission

The purpose of this paper is to present recent advances made by EUROSTAT, the statistical office of the European Union, in the area of GDP nowcasting. In addition to recently implemented T+30 methodology for GDP flash estimates, EUROSTAT has implemented its own modeling approach for nowcasting GDP as part of its work on the Principal European Economic Indicators (PEEIs).

EUROSTAT's GDP nowcasting methodology is based on the use of two nowcasting models, namely an Error Correction Models (ECM) time-series model (bridge model) and a MIDAS (Mixed-Data Sampling) time series model. The MIDAS model is a more general ECM that takes into account the mixed (quarterly/monthly) frequency data which is advantages when nowcasting GDP. Euro Area quarterly GDP data series have been converted to monthly frequency using Chow Lin disaggregation method. Monthly data for all indicators have been used from January 1999 to February 2017, using GDP, retail sales and exports using industrial production, construction output, exports, unemployment rate, Euro-dollar exchange rate as input. In a complementary approach, EUROSTAT has also used combined 'hard' with 'soft' data. In order to run the different models, the four coincident series (industrial production, consumption in manufactured goods, exports and MSCI Price Index) have been forcasted first using a conventional ARIMA approach.

The methodological approach implemented to date has so far been implemented for the Euro area and one of its member states, France. The results obtained are promising. A cross-comparison of different modelling results as well as of the results from the official T+30 submission for EU/EA member states has been carried out. Results from the proposed MIDAS model mach the actual values most closely. EUROSTAT will continue its work in this area to investigate if the MIDAS approach is to be preferred to the 'pure' ECM approach.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Time to peak concentration for bioequivalence: Analysis to support commercial claim

Michelle Foster
Reckitt Benckiser

Nurofen is a global brand for Reckitt Benckiser (RB) and is a popular over-the-counter drug for effective pain relief. RB conducts many relative bioavailability studies to help support RB's on-going global registrations and claim substantiations. Area under the curve and maximum concentration are commonly used as endpoints to conclude bioequivalence. Also, time to reach maximum concentration (Tmax) may be used as a pharmacokinetic endpoint to provide additional insight with respect to claims.

The non-parametric Wilcoxon matched pairs test is commonly used to analyse Tmax in different formulations of a product. However, this test limits itself to partial use of the survival curve. Specifically, the median Tmax focuses on only one point on the Kaplan Meier survival curve for each treatment arm and provide an absolute measure of similarity in pharmacokinetic profiles with no indication on efficacy.

In attempt to optimise information, we use proportional hazards regression to analyse data from a study by RB which compared the speed of absorption of three formulations of Nurofen: Ibuprofen lysinate powder (test product), standard Ibuprofen (reference) and Ibuprofen lysine tablets (comparator).

The results suggest that there is difference in Tmax for both test and comparator compared to the reference. Specifically, the likelihood of reaching maximum concentration is 11 times and 5 times faster for those on test and comparator respectively compared to the reference. In summary, claims of superiority of the test and the comparator over the reference are better substantiated by reporting hazard ratios together with median difference in Tmax.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Bayesian Inference for the Coalescent model using Sequential Monte Carlo samplers with across-model methods

Richard Culliford, Richard Everitt, Daniel Wilson
University of Reading

We present a Sequential Monte Carlo (SMC) framework, which we term 'transformation Sequential Monte Carlo', with a key focus in the field of population genetics. The proposed algorithm is aimed for the special case where the total number of parameters to be inferred in their corresponding joint Bayes posterior distribution increases when an additional observation is added. This differs in comparison to a sequence of sequential distributions still retaining a common parameter space, when the difference between distributions is the sample size. Therefore we transition from a previously inferred model to a model of differing dimensional space using across-model methodology. A gradual annealing transition between the two models is incorporated in the algorithm to compensate for the scenario where a suitable transformation function is hard to construct or if there is uncertainty regarding how appropriate said function is in exploring the parameter space. We present results from performing Bayes inference on ancestral trees, under the coalescent model, starting from a smaller subset of sequence observations towards a larger set of full sequences by sequentially grafting each sequence one at a time onto the tree. The core difference from the majority of algorithms applied to this application is how it builds upon an existing tree, therefore it is possible to allow for any removal of tip nodes or a change of genealogy assumptions without the need to initiate a new Markov Chain under different model conditions.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Laspeyres-type what?! a European notion of Laspeyres, Lowe and Young

Jens Mehrhoff
European Commission

The "Laspeyres-type index" label is imprecise; what is meant in most cases is actually the Lowe index but also the Young index can be referred to. Taking the legal act for the European Harmonised Index of Consumer on Prices (HICP) as a starting point, this paper elaborates how Laspeyres-type indices are calculated in the EU.

The target formula for the HICP would be the Laspeyres index, which however cannot be calculated in real time due to the unavailability of weights from the previous year (t–1) at the beginning of the next year. Hence, data from the year before the previous year (t–2) form the basis for the estimation of weights that are as representative as possible for consumers' expenditure patterns in the target weight reference period. Furthermore, the estimated expenditure shares for the year t–1 have to be adjusted (or price-updated) to the price reference period which for the HICP is December of the year t–1.

In addition to the practical issues as applied in national statistical institutes, the paper also investigates under which circumstances the Lowe or the Young index will be the better approximation to the true Laspeyres index. Though the HICP is designed to assess price stability and is not intended to be a cost-of-living index, it turns out that the answer to this question is related to consumer substitution behaviour.

Finally, the paper gives an outlook to the forthcoming Methodological Manual of the HICP, to be published in early 2018. The manual explains preferred methods, illustrated by examples of good practice.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

A stratified randomized response techniques for collecting information on sensitive characteristics

Olaniyi Mathew Olayiwola, Adebisi Agnes Olayiwola
Federal University of Agriculture, Abneokuta, Nigeria

Randomized Response Techniques (RRT) enhances protection of respondent's identity and allows collection of reliable information on sensitive and stigmatized characteristics. There is a dearth of information about the RRT for heterogeneous population. This study was therefore designed to develop a Stratified Randomized Response Technique (SRRT) with three randomization devices. A simple random sampling scheme was used to select 500 students which comprised of 332 males and 168 females. These students were stratified into two strata using gender as stratifying factor and proportional allocation procedure was used to select 199 male students and 101 female students randomly. An interviewer administered questionnaire was used to collect data on demographic characteristics and related sensitive questions The selected students were given three randomization devices (coin, dice and deck of cards) out of which one was randomly selected by the respondents. When a coin is picked and tossed with head as the outcome, a deck of cards is picked and shuffled with heart, spade and diamond as outcomes and a die is picked and tossed with outcome 1, 2, 3 and 4, then the respondents were asked to answer the sensitive questions secretly and truthfully. An estimator was proposed for estimating the proportion of students with sensitive behaviour and its statistical properties were examined. The proportion of male students who had contracted sexually transmitted disease, involved in rape, abortion, cultism, murder, stealing and examination malpractice is higher than that for the females. The proposed estimator showed that the variance of proportion of female students involved in the sensitive and stigmatized characteristics is higher. The proposed estimator showed that the variance of proportion for both male and female students involved in the sensitive and stigmatized characteristics is lower. Hence, the SRRT estimator is more efficient

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Tree-based Sampling Algorithms for Particle Smoothing in a Hidden Markov Model

Dong Ding, Axel Gandy
Imperial College London

We provide a new strategy built on the divide-and-conquer approach (Lindsten et al., 2017, Journal of Computational and Graphical Statistics) to investigate the smoothing problem in a hidden Markov model. The proposed tree-based sampling algorithm decomposes a hidden Markov model into sub-models with a binary tree structure and gradually merges and resamples the particles generated from the sub-models towards the complete model in the root of the tree using importance sampling. The key question we address is the choice of the proposals and target distributions of the sub-models in the tree. In particular, we propose a type of intermediate target distributions which leave the marginals invariant. We further implement iterative runs of the algorithm to update the proposal and target distributions in each level of the tree. In the simulation studies, the proposed tree-based smoothing algorithm works comparably well regarding the mean square error and also produces far more particles than the conventional smoothing algorithms given the same computational effort.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Time Series Modelling and Forecasting of Hospital Overcrowding in Ireland

Jean Abi Rizk, Cathal Walsh
University of Limerick

Prize winner 2017 Research Student Conference

According to the daily trolley count by the Irish Nurses and Midwives Organisation (INMO), the overcrowding crisis in Irish hospitals has reached a new record at the start of 2017 with more than 600 patients on trolleys. This work presents a time series approach to model the INMO data and to present future predictions of the overcrowding in Irish hospitals. The structure of the data exhibits short and long seasonal patterns along with a moving event that occurs at the start of every year. While the widely used ARIMA models fail to capture multiple seasonality with long seasonal periods, we present a time series modelling approach that is a combination of a Fourier series to model the long seasonal pattern and a seasonal ARIMA process to model the short-term dynamics. The model shows reasonable forecasts that approximately match the actual 2017 data showing that the high records of 2017 could have been predicted in advance. The model predictions may help the healthcare managers to take alternative courses of action and give them the opportunity to plan better for the future.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Utilisation and impact of community-based antiplatelet therapy on outcomes in patients following Acute Coronary Syndrome in Scotland during 2012-2014

Grant Wyper, Samantha Alvarez-Madrazo, Kim Kavanagh, Martin Denvir, Marion Bennie
NHS National Services Scotland

Introduction

Acute Coronary Syndrome (ACS) is a life threatening condition that occurs when blood flow to the heart is reduced due to a blockage of a coronary artery. A partial occlusion is defined as a Non ST-segment elevation myocardial infarction (NSTEMI) or Unstable Angina (UA). The most severe presentation is a complete blockage, characterised as ST-segment elevation myocardial infarction (STEMI). Adherence to antiplatelet therapy is essential for the secondary prevention of atherothrombotic events. The aim is to assess the utilisation of traditional and novel antiplatelets following ACS and investigate the impact of community-based treatment on efficacy and safety outcomes.

Methods

Secondary care diagnoses were used to identify a retrospective cohort and to retrieve historic and subsequent inpatient hospital diagnoses to allow the identification of prior and prospective healthcare events within the sampling frame. Record linkage using a common patient-identifier was used to obtain community dispensed drug therapies. Information related to the dosage, unit and frequency of therapy was extracted from free-text dosage instructions. A survival analysis will be carried out to assess the differences between novel and traditional therapies in relation to clinically relevant cardiovascular, mortality and safety events, taking into considering any disparities in adherence to therapy by patients.

Results

29,708 patients presented with ACS (50.6% NSTEMI; 28.9% STEMI; 8.9% UA and 11.6% unspecified). STEMI patients were the younger (Mean=63.9 years; SD=13.5) than NSTEMI patients (Mean=69.0 years; SD=13.5) and were the group with the largest percentage of males (67.6%). There was a significant difference in in-hospital mortality between clinical presentations ($p < 0.01$).

Conclusion

These findings will provide real-world evidence on the utilisation and impact of different antiplatelet regimens. Through considering the real-world varying constraints of individuals with multimorbidity, adherence and persistence to therapy, these findings will provide significant insights into the clinical impact of these medicines in Scotland.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Relative Importance of Functional Predictors in Scalar on Function Regression

Alkeos Tsokos, Ioannis Kosmidis
University College London

We explore the relative importance of functional predictors in scalar on function regression, where importance is defined in terms of percentage of variance explained in the response. The properties of inferential procedures for relative importance are studied in comprehensive simulation studies, and the proposed methodology is applied to a real data set. The inferential procedures depend heavily on the variance-covariance matrix of the estimators of the regression parameters. We assess the performance of various estimators of that matrix, including approaches that account for the variability associated with choosing the tuning parameters in the roughness penalty to the log-likelihood.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Transforming the Practical Driving Test

Caroline Wallbank, Sritika Chowdhury, Jo Hammond, Lauren Durrell, Neale Kinnear, Su Buttress, Shaun Helman
TRL

Practical driving tests are designed to ensure new and novice drivers show a certain level of competence before being allowed to drive unsupervised. Therefore, the components of the driving test are important to test the skills and ensure safety of these drivers.

The Transport Research Laboratory was recently involved in trialling some changes to the current practical driving test with the Driver and Vehicle Standards Agency (DVSA). The purpose of the proposed changes to the test was to improve the way novice drivers prepare for ‘real world driving’, thus having an effect on their safety and attitudes.

Both quantitative and qualitative methods were used to obtain data from over 4,300 drivers and 860 driving instructors to examine what impact the trial test has on preparation, attitudes to risk, post-test driving, and on the number of self-reported collisions in the first six months of driving after passing the practical driving test. Longitudinal research methods were used to collect data and participants were randomly assigned to take the current or trial test.

Generalised Linear Modelling was used to examine the relationship between collisions and a number of key variables including the amount and type of learning pre-test, responses to attitudinal questions and exposure to driving post-test. Results from the statistical modelling showed that some variables related to pre-test and post-test exposure were associated with collision risk. There was some evidence that the new test may be a good basis for improving pre-test driving in the future. Based on the results, the DVSA has announced that from December 2017 all novice drivers will now take the ‘new’ test, which includes the trial components examined as part of this research study.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Surrogate evaluation using information theory: extension to the case of ordinal outcomes

Hannah Ensor, Christopher Weir
Biomathematics and Statistics Scotland (BioSS)

Background

Surrogates are measures used to predict treatment effect on true outcomes. Replacing true outcomes with valid surrogates can reduce the length and size of a clinical trial.

Objectives

A leading approach to evaluate potential surrogates is based on information theory (Alonso & Molenberghs, 2007). We extended this approach to the case of a binary surrogate and ordinal true outcome (binary-ordinal setting), evaluating this extension through a large simulation study.

Methods

In the presence of multiple trials, the two-stage information theory approach assesses the amount of information on the treatment effects on the true outcome provided through adjusting for the treatment effects on the surrogate. In the binary-ordinal setting we do this using a likelihood reduction factor based on proportional odds models. We assessed this approach under simulation study investigating various: numbers of patients per trial; numbers of trials; surrogacy strengths; and adherence to the proportional odds assumption.

Results

In the simulation study the approach generally estimated surrogacy strength well across scenarios. However, bias was imposed due to the use of a pragmatic two-stage approach.

Conclusions

This work will provide researchers in clinical areas where ordinal outcomes are utilised with a means of evaluating surrogates.

Alonso, A., & Molenberghs, G. (2007). Surrogate marker evaluation from an information theory perspective. *Biometrics*, *63*(1)

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

A solution for the polarimetric statistical detection problem using a quaternion RKHS

Antonia Oya, Jesús Navarro-Moreno, Juan Carlos Ruiz-Molina
University of Jaén

The key to the resolution of many statistical signal processing problems is a suitable method for characterizing random processes. The selection of an appropriate one usually facilitates the problem we try to solve. In this paper, we are interested in using a quaternion reproducing kernel Hilbert space (RKHS) approach to deal with the detection of a polarimetric signal disturbed by additive Gaussian noise. There exist several models to represent the polarization state of an electromagnetic wave; but, Jones vector representation is one of the most widely used models in Physics. Specifically, a polarimetric signal can be expressed by a two-dimensional complex vector in which the possible statistical relationship between both components is the polarization information. Recently, an equivalent representation of the polarization state based on the quaternion widely-linear model has been applied in several fields. It is the Cayley-Dickson representation of quaternions which allows us to express the Jones vector as a quaternion. Moreover, the quaternion widely-linear processing which requires the augmented statistics to be considered is the suitable statistical processing for quaternion signals.

Use of the RKHS approach for the problem of signal detection in Gaussian noise is a classical method in signal processing literature. A coordinate-free representation of the augmented quaternion noise based on the RKHS associated with its augmented correlation function will allows us to obtain a unified solution to the signal detection problem considered. In this work, we applied the quaternion RKHS approach to derive an expression for the log-likelihood detector in terms of inner products in the RKHS corresponding to the augmented quaternion noise. The main property that makes the RKHS representation useful is that the suggested expression for the log-likelihood ratio unifies a variety of formulas for the optimum detection statistic (for instance, in terms of series expansions, solutions to integral equations, etc.).

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

ONS's Progress towards an Administrative Data Census: 2017

Abigail Higgins, Jaspreet Gakhal, Matthew Roberts
Office for National Statistics

Prize winner 2017 Young Statisticians Meeting

The Office for National Statistics are looking beyond the next census in 2021 to the possibility of putting together a census based on linking together administrative data held by the Government. We're calling this an Administrative Data Census. This approach could offer benefits such as the production of more timely, frequent data in addition to new opportunities, at a lower cost.

Our main goal is to produce as many census-type statistics as possible. We'll do this by combining administrative and survey data, and comparing them with the outputs from the 2021 Census.

Every year we publish an Annual Assessment report to demonstrate ONS's progress towards moving to an Administrative Data Census post 2021 and our second assessment was published in May 2017.

This assessment is conducted against a set of high level criteria which need to be in place before we can move to an Administrative Data Census:

- Rapid access to existing and new data sources
- The ability to link data efficiently and accurately
- Methods to produce statistical outputs of sufficient quality that meet priority information needs of users
- Acceptability to stakeholders (users, suppliers, public and Parliament)
- Delivering value for money

Each assessment provides an update on our current position and where we expect to be in 2023.

This poster describes our progress towards an Administrative Data Census and summarizes how the assessment demonstrates this progress.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Comparison of the Genetic Algorithm and Incremental Optimisation routines for a Bayesian inversion optimal network problem

Alecia Nickless
Nuffield Department of Primary Care Health Sciences, University of Oxford

The estimation of fluxes of trace gases by means of Bayesian inversion requires a network of atmospheric monitoring sites where the concentrations of these gases are measured. Different solutions for the optimal placement of a five-member atmospheric monitoring network in South Africa for the measurement of CO_2 are compared between various configurations of a genetic algorithm (GA) against an incremental optimisation (IO) routine. The cost function for the optimisation is derived from the posterior covariance matrix of the estimated fluxes. This is calculated from the Bayesian inversion solution for estimates of CO_2 sources in space and time under a given observation network, and does not require the observed concentrations. The posterior covariance matrix, when compared to the prior, conveys the uncertainty reduction in the flux estimates achieved by a particular configuration of measurement sites.

GA's are a class of stochastic optimisation procedures which consider all parameters simultaneously, where the parameters solved for here are the locations of the measurement stations. The “population size” and the number of iterations need to be specified for a GA, where computational cost increases with each of these. The IO routine computes the cost function from the set of possible locations incrementally, first selecting the station which produces the greatest reduction in uncertainty in a one-member network, followed by the station which, when added to the network, results in the largest additional uncertainty reduction, and so on. This method is more efficient in terms of computational resources, but may not always find the global optimum in a multiple-parameter problem. I introduce a statistic to compare how spatially similar the placement of stations is between two different networks, and show that there is a small advantage in the use of the GA over the IO routine, but this comes at a large computational cost.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Medical Cost Prediction Using Whole Salford IGR/Diabetes Patients by Applying a Bayesian Network

Shuntaro Yui, Toshinori Miyoshi, Takanobu Osaki, Hideyuki Ban, Norman Stein, Sheila Mccorkindale, Martin Gibson
Hitachi Ltd

We have been establishing a cost modelling framework in order to demonstrate the degree to which delaying or preventing the onset of type2 diabetes could be economically effective. Last year, we have presented preliminary cost modelling results using the Salford integrated record system across both primary and secondary care which allowed the opportunity to capture detailed analysis across both spectrums of care at the same time. However, the cohort data was incomplete because of incomplete cohort criteria. In this poster, we present economic modelling using new cohort data by applying a Bayesian Network.

In new cohort data, we newly included outpatient, A&E data, IGR patients who don’t have IGR disease code but whose HbA1c/blood glucose level met criteria, and IGR/diabetes patients whose secondary diagnosis are IGR/diabetes. Meanwhile, we deleted the patients whose disease code was lifestyle disease only such as hypertension and obesity in order to understand disease progression from IGR to diabetes. There are some research using a Markov model, however, our approach was a Bayesian network based method, because 1) Our method interpolates significant missing data values because the new model can describe the complex relationship between each indication such as test results and disease status. 2) It also introduces a data-driven approach to model construction compared to the Markov model based method.

Experimental results using Salford IGR/diabetes data (11869 patients) showed that the amount of data available was twice larger than Markov model based method because of missing data inclusion and less leverage of disease expansion, whilst predicted medical cost using available data in Markov model was £761, that using whole data was £1000 in which prediction error achieved less than 1%. It demonstrates that the proposed method can achieve future diabetes medical cost without complicated efforts by interpolation of missing data values and semi-automatic model construction.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Exploring missingness mechanisms for which complete case and multiple imputation analyses produce similar regression coefficient estimates

Finbarr Leacy
Royal College of Surgeons in Ireland

Despite making different assumptions about the underlying missing data mechanism, complete case and multiple imputation analyses can sometimes produce similar regression coefficient estimates. This apparent agreement can mislead applied researchers who, without considering whether either set of assumptions is plausible for the data at hand, may erroneously conclude that such estimates are always unbiased.

Drawing on a selection of recent work in the missing data literature [2-7] and using simulated data with multivariate missingness, we present a range of simple examples where complete case and multiple imputation analyses produce similar linear and logistic regression coefficient estimates, including cases where both sets of estimates are equally biased. These examples reiterate the importance of carefully examining the missingness mechanism prior to analysis.

References

[1] Cornish RP, Tilling K, Boyd A, Davies A, Macleod J. *Int J Epidemiol.* 2015; 44(3):937-45.

[2] White IR, Carlin JB. *Stat Med.* 2010; 29(28):2920-2931.

[3] Little RJ, Zhang N. *J R Stat Soc Ser C.* 2011; 60(4):591–605.

[4] Bartlett JW, Carpenter JR, Tilling K, Vansteelandt S. *Biostatistics.* 2014; 15(4):719–730.

[5] Bartlett JW, Harel O, Carpenter JR. *Am J Epidemiol.* 2015; 182(8):730–736.

[6] Galati JC, Seaton KA. *Stat Methods Med Res.* 2016; 25(4): 1527-34.

[7] Little RJ, Rubin DB, Zangeneh S. *J Am Stat Assoc.* 2017; 112(517): 314-320.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Convex Relaxation reformulation of the Goodness-Of-Fit test for copulas

Victory Idowu
London School of Economics

Copulas have increasingly become the main tool for modelling dependence between random variables. Increasingly there has been a need for a Goodness-Of-Fit measure of the quality of a copula to a given data set. To investigate this problem, the following statistical test setup is used: a hypothesis test with a null hypothesis providing a member from a relevant copula family and an alternative. There is also a need to distinguish between copula families and possible parameterisation which is not covered by the current statistical test setup.

We extend the principles behind the Goodness-Of-Fit to present a methodology which represents the Goodness-Of-Fit problem using Convex Relaxation.

Convex Relaxation is an approach used in optimization theory which reduces the complexity of constraints in the programme by replacing them with convex functions. In doing so, one can easily distinguish between copulas and solve the algorithm within computational time. In addition, information criteria can be derived; which can be used to formulate plausible parameters for a copula from a certain family.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Projects in the ONS 'Economic Statistics Centre of Excellence'

Stuart McIntyre, Gary Koop, James Mitchell, Katerina Lisenkova, Graeme Roy
University of Strathclyde

This poster will highlight the work of two projects taking place as part of the ONS Economic Statistics Centre of Excellent, summarized as:

Gary Koop (Strathclyde) James Mitchell (WBS) and Stuart McIntyre (Strathclyde)

Regional nowcasting in the UK: The project has four aims. Firstly, to produce and disseminate timely model-based quarterly regional estimates of nominal GVA to the same timetable as the UK's first estimates of quarterly GVA for the UK as a whole. Secondly, to produce historical quarterly estimates of regional GVA, if feasible, at greater levels of regional and sectoral levels of disaggregation using Big Data econometric methods. Thirdly, to produce real or volume GVA estimates, using the ONS's "experimental" real regional GVA data. To explore the use of alternative ways of estimating regional GVA deflators. Finally, to explore, jointly with ONS staff, the possible use of underlying micro-level and administrative data to produce model-free (or less model dependent) quarterly regional output data.

Steve Gibbons (LSE) Katerina Lisenkova (Strathclyde), Graeme Roy (Strathclyde) and Kim Swales (Strathclyde).

Improving the quality of regional economic indicators: it is clear there is an increasing focus and attention on regional economic performance and the devolution of powers in the UK. Scotland is at the forefront of devolution and we believe provides a useful case study to begin developing new regional indicators.We plan to undertake two projects, to take forward a stream of work to improve the quality and usefulness of regional economic statistics. The first of these focuses on improving inter-regional trade flows data. The second area of work, which would take place beyond the project end date shown, is to further improve the quality and robustness of regional fiscal data

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Attributing changes in the distribution of species abundance to weather variables using the example of British breeding birds

Cornelia Oedekoven, David A. Elston, Philip J. Harrison, Mark J. Brewer, Stephen T. Buckland, Alison Johnston, James Pearce-Higgins
University of St Andrews

Modelling spatio-temporal changes in species abundance and attributing those changes to potential drivers such as climate, is an important but difficult problem. The standard approach for incorporating climatic variables into such models is to include each weather variable as a single covariate via a low-order polynomial or smoother in an additive model. This, however, confounds the spatial and temporal effects of the covariates.

We developed a novel approach to distinguish between three types of change in any particular weather covariate. We decomposed the weather covariate into three new covariates by separating out temporal variation in weather (averaging over space), spatial variation in weather (averaging over years) and a space-time anomaly term (residual variation) which were each fitted separately in the models. We illustrate the approach using generalized additive models applied to count data for five species from the UK's Breeding Bird Survey, 1994-2013. The weather covariates considered were temperatures during the preceding winter and temperatures and rainfall during the preceding breeding season. We compare models that include these covariates with models including decomposed components of the same covariates, considering both linear and smooth relationships.

Lowest QAIC values were always associated with a decomposed weather covariate model. Different relationships between counts and the three new covariates provided strong evidence that the effects of changes in covariate values depended on whether changes took place in space, time, or in the space-time anomaly. These results promote caution in predicting species distribution and abundance in future climate, based on relationships that are largely determined by environmental variation over space.

Our methods estimate the effect of temporal changes in weather, whilst accounting for spatial effects of long-term climate, improving inference on overall and/or localised effects of climate change. Our methods represent an important advance by eliminating the confounding issue often inherent in large-scale data sets.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Detection of change points in count data

Taghreed Mohammed Jawa, David Young, Chris Robertson
University of Strathclyde

In Scotland, healthcare associated infection (HAI) is a major factor of patient morbidity and mortality, especially methicillin-resistant staphylococcus aureus (MRSA) bacteraemia. It is also associated with increasing costs so Health Protection Scotland established and improved healthcare interventions to control infection and avoid HAIs. Some interventions took place in Scotland to tackle the rates of HAIs and the infection rates subsequently decreased. It is of interest in epidemiology to know when changes occur in order to identify healthcare interventions associated with these changes. Several statistical methods are investigated in this research to detect the time when rates of MRSA bacteraemia changed and to determine which associated interventions may have impacted the rates.

Change points are estimated from polynomial generalized linear model (GLM) which account for seasonality and confidence intervals for the change points are constructed using a bootstrap method. Segmented regression is also used to detect the change points at times when specific interventions took place. Joinpoint analysis looks for potential change points at each time point in the data. The joinpoint model is adjusted by adding a seasonal effect to account for additional variability in the rates. Confidence intervals for the joinpoints are constructed using bootstrapping. Change points are also estimated using the spline function in the generalized additive model (GAM) and bootstrap confidence intervals for the change points are constructed.

Polynomial Poisson regression and spline GAM methods detected the change points when rates of MRSA bacteraemia decreased during 2005- 2006 while segmented regression and joinpoint analysis found the change points during 2006- 2007. All methods were found to detect similar change points. Segmented regression is used to detect the actual point when an intervention took place. Polynomial GLM, spline GAM and joinpoint analysis models are useful when the impact of an intervention is observed after a period of time.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Modelling the probability of liquidations in UK manufacturing industries with dynamic Bayesian networks

David Purves, Lesley Walls, Quigley John
University of Strathclyde

Objectives

The ability to forecast whether a company will cease trading within a particular period of time can be crucial to the strategic planning of partner companies. We constructed a dynamic Bayesian network (DBN) using a mixed-effects methodology to model time-series data of financial returns and associated liquidation status. Estimated models were used to generate predictions for various company risk profiles using the conditional distribution and value-of-information techniques illustrated.

Data

The HMRC Datalab provided access to end-of-year financial statements and the liquidation status of UK listed companies, for the years 2000 to 2012. Each company self-reported an industry classification which were subsequently grouped within broader industry classes using an internal HMRC classification number. Using this database, a range of financial ratios were derived, and together with company age, gross domestic profit, and indicators relating to late submission of accounts, were used in estimating a DBN.

Methods

To estimate the DBN we assumed that a global structure could be used to describe the associations between the variables for all companies. We used mixed-effects models so that dependencies within industry subtypes were appropriately accounted for, while allowing for predictions to be produced for company's with few records. Two DBN's were estimated using model averaging, on a dataset with observations with any missing data removed, and by using indicators when data were missing. The DBN's were parameterised using the conditional distribution, allowing for industry subtype level trends.

Results
(Full results to be released)

For each DBN a range of predictive measures were generated in-sample and using a hold-out sample. For predicting liquidation status, we found that increased bank savings, and retained profit and loss had a protective effect, and increased total liabilities the reverse (variables were normalised by total assets).

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Bayesian inference for bivariate copulas with additive models for dependence, marginal location, scale and shape: an application in paediatric ophthalmology

Mario Cortina Borja, Julian Stander, Luciana Dalla Valle, Charlotte Taglioni, Angie Wade, Brunero Liseo
University College London

Motivated by data on visual acuity from a large sample of children aged between 3 and 8 years, we propose bivariate copula models with dependence parameters, and marginal densities with location, scale and shape parameters that depend on a covariate through additive models. We perform inference about the unknown quantities of our model in the Bayesian framework, using a Markov chain Monte Carlo algorithm, the proposal steps of which take account of parameter constraints. We model the marginals the copula function with the four-parameter sinh-arcsinh distribution as defined within the class of generalised linear additive models for location, scale and shape (gamlss).

We apply this model to paediatric ophthalmic data to understand the process which causes changes in visual acuity with respect to age. In particular, we are interested in age-related changes in the dependence parameter of the copula model. We analyse predictive distributions to identify children with unusual sight characteristics, and discuss extensions to our model.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Effectiveness of Country Durham Wellbeing for Life intervention in improving the health of deprived communities

Nasima Akhter, Shelina Visram, Adetayo Kasim
Durham University

Introduction: The County Durham Wellbeing for Life (WFL) service, an integrated service aimed at improving health and wellbeing of deprived communities, includes one-to-one support from a Health Trainer (HT) to deal with health issues through developing a personalised care plan and delivering targeted advice on behaviour change. A nationally implemented Data Collection and Reporting System (DCRS) captures information on implementation processes and outcomes of HT services. This provides an opportunity to evaluate the effectiveness of HT component for the WFL service in Durham.

Objective: This study evaluated success of the WFL intervention in terms of improved physical and mental health and wellbeing among service users.

Method: Anonymised, individual-level DCRS data for clients who had completed the WFL one-to-one intervention between June 2015 and January 2017 were extracted and analysed. Physical health outcomes (EQ5D5L and EQ5D visual analogue scale (EQ5DVAS)) and mental health outcome (Short Warwick-Edinburgh Mental Wellbeing Scale (SWEMWBS)) were analysed in SAS 9.4. Mixed effect linear models that included age, gender, ethnicity, deprivation status, primary issue, providers, area, and significant interaction for deprivation and time were used to account for correlation between repeated measures per individuals.

Results: Compared to baseline, improvements were evident at completion of the intervention: 4.1 points (Confidence Intervals: 3.1, 5.1); 12.3 points (10.9, 13.6); 10.2 points (9.0, 11.4) points for EQ5D5L, EQ5DVAS and SWEMWBS, respectively. Improvements sustained with lower magnitude: 1.7 (0.0, 3.4), 2.0 (-0.3, 4.4) points for EQ5D; 8.4 (6.2, 10.7), 10.0 (6.4, 13.5) points for EQ5DVAS; and 9.0 (7.0, 11.0) and 7.0 points (3.9, 10.2) for SWEMWBS at six and 12 months, respectively.

Conclusion: The one-to-one HT component of the WFL service had positive health outcome at completion and during follow up period.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Detection of Safety Signals in Randomised Controlled Trials using Groupings

Raymond Carragher
University of Strathclyde

The occurrence, severity, and duration of patient adverse events are routinely recorded during randomised controlled clinical trials. This data may be used by a trial's Data Monitoring Committee to make decisions regarding the safety of treatments and in some cases may lead to withdrawal of treatment if real safety issues are detected. There are many different types of adverse event and the statistical analysis of this data, particularly with regard to hypothesis testing, must take into account potential multiple comparison issues. Unadjusted tests may lead to large numbers of false positive results but simple adjustments are generally too conservative. In addition, the anticipated effect sizes of adverse events in clinical trials are generally small and consequently the power to detect such effects is low.

Recently a variety of classical and Bayesian methods have been proposed to address this problem. These methods use possible relationships or groupings of adverse events in their statistical analyses. We illustrate and compare a number of these recent approaches, and investigate if the use of a common underlying model which involves groupings of adverse events, by body-system or System Organ Class, is useful in detecting adverse events associated with treatments. All of the group methods considered correctly flag more adverse event effects than the standard error controlling approaches, such as the Benjamini-Hochberg procedure or Bonferroni correction, for this type of data. However, none of the methods take into account event timings or total exposure time for patients. In order to address the desire for early detection of safety issues in a clinical trial a number of the Bayesian methods are extended to analyse the accumulation of adverse events as the trial progresses, taking into account event timing and patient time in study. These methods are suitable for use at interim trial (safety) analyses.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

A systematic root cause analysis into the increase in Escherichia coli bacteraemia in Wales over the last 10 years

Jiao Song, Angharad Walters, Damon Berridge, Ronan Lyons, Ashley Akbari, Mari Morgan, Maggie Heginbothom
School of Medicine, Swansea University

Background

Public Health Wales have been requested to undertake investigation into the rise of *E. coli* bacteraemia in Wales over the last 10 years by the Chief Medical Officer for Wales.

Data

Anonymised blood microbiology culture data reported between 2005 and 2011 are included in the Secure Anonymised Information Linkage (SAIL) databank which holds anonymised, routinely collected administrative data. *E. coli* bacteraemia cases have been linked with Welsh Demographic Service (WDS) data, Patient Episode Database for Wales (PEDW), Welsh general practice data and Welsh Index of Multiple Deprivation (WIMD) via anonymous linkage field (ALF) to flag relative risk factors. These factors include age, comorbidity score, operation history and so on.

Methods

A case-control study was conducted. All potential controls were randomly selected from WDS. Three different methods were used to identify controls: 1) cases and controls had a Welsh address on the date the *E. coli* blood sample of the case was received; 2) method 1 was extended to also match on age and gender; 3) method 2 was extended by additionally matching on GP practice. Logistic regression and conditional logistic regression modelling techniques were used to identify risk factors for developing *E. coli* bacteraemia. A sensitivity analysis was performed to check the robustness of results to a change in the method of selecting controls.

Results

Method 1 showed that urine infection (OR 21.53, 95% CI 19.57-23.69), hospital antibiotics prescription (OR 16.46, 95% CI 15.05-18.02) and high comorbidity score (OR 16.02, 95% CI 14.66-17.50) are the risk factors with the highest odds ratios. Other two methods obtained similar results. Standard errors were adjusted for clustering within case-control groups.

Conclusion

Identifying patients at highest risk can be improved after determining the factors associated with development of *E. coli* bacteraemia. Preventive interventions can be introduced if these risk factors are modifiable.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Hierarchical Bayesian Models for Road Traffic Accidents

Zainab Al-Kaabawi, Yinghui Wei, Rana Moyeed, Malgorzata Wojtys
Plymouth University

Injuries from road traffic accidents are a public health problem worldwide, being one of the leading causes of death according to the World Health Organisation. We aim to estimate the intensity function of accidents and study its pattern throughout the UK motorway network. We estimate the road-specific intensity function by modelling the point pattern of the accident data using a homogeneous Poisson process. The estimated intensity across roads is then combined to obtain the overall intensity function which provides an estimate of the between-road heterogeneity. We adopt a Bayesian approach to inference and use MCMC methods for extracting the posterior distributions of the unknown parameters. A sensitivity analysis is conducted by using different priors. We evaluate the performance of the proposed methods using a simulation study.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

The modality of multivariate skew normal mixture

Bader Alruwaili
Glasgow University

Finite mixtures are a flexible and powerful tool that can be used for univariate and multivariate distributions, and a wide range of research analysis has been conducted based on the multivariate normal mixture and multivariate of a t-mixture. Determining the number of modes is an important activity that, in turn, allows one to determine the number of homogeneous groups in a population. Existing research in this area have attempted to identify the upper bound of the number of modes for a normal multivariate mixture. Our work currently being carried out relates to the study of the modality of the skew normal distribution in the univariate and multivariate cases. For the skew normal distribution, the aims are associated with studying the modality of the skew normal distribution and providing the ridgeline, the ridgeline elevation function, the Π function, and the curvature function, and this will be conducive to an exploration of the number and location of mode when mixing the two components of skew normal distribution. The subsequent objective is to apply these results to the application of real world data sets, such as flow cytometry data.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

A predictive model for HIV using a semi-parametric spline approach

Amos Chinomona, Henry Mwambi
Rhodes University

The generalized additive models (GAMs), extensions of the generalized linear models (GLMs), enable exploring the non-linear dependence of a response variable on predictor(s) variable(s) in a non-parametric or a semi-parametric framework. GAMs are often used when there is no *a priori* reason for determining a particular response function in a regression setting and allow the data to “speak for themselves”. This is achieved via the use of smoothing functions. The objectives of the research is to construct a semi-parametric model for HIV that is flexible in allowing the data to determine its form and nature.
A semi-parametric logistic GAM for HIV on demographic, socio-economic and behavioural variables using population-based 2010-11 Zimbabwe Demographic and Health Surveys (2010-11ZDHS) data is fitted. The dependence of HIV on the non-parametric smooth function of the variable age, as a continuous covariate, and parametrically on the other demographic and socio-economic factors was investigated. The results were compared with the results of an equivalent ordinary logistic GLM from a likelihood perspective.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Let’s meet in the middle: facilitating access to administrative data in the UK

Rowan Lawrance, Ilse Verwulgen, Elaine Mackey
Administrative Data Research Network

The Administrative Data Research Network (ADRN) facilitates access to de-identified administrative data for researchers. Under a complex and dynamic data sharing legal framework in the UK, the Network is a partnership of UK Universities, government departments, national statistical authorities, funders and research centres and it aims to deliver a service enabling secure and lawful access to de-identified linked administrative data to researchers.

As one of the ‘front doors’ to the ADRN, the Administrative Data Service is liaising with data owners, researchers and experts in data linkage and data governance to facilitate access to administrative data. In addition to providing guidance on processes and an infrastructure addressing some of the concerns on information governance and data security through dedicated ‘secure environments’ as points of access. Quite often, we find ourselves in the ‘middle’ of these discussions, as we negotiate access and translate requirements and repurpose documentation to ensure the project resonates with a variety of agendas and priorities.

The poster will provide an overview of recent work in the area and how we have dealt with challenges up to now. We will summarise work done in trying to streamline application processes for different data providers in different data domains in the UK (e.g. education, health, crime, benefits and labour market). We will talk about how ADRN has been working alongside government departments to design and implement streamlined approaches to administrative data access in the UK and how we are supporting researchers when they apply to access administrative data for their research in the areas of ethics, consent, legal pathways to access, methodology and data availability. And how it’s not just about data meeting in the middle, it’s primarily about people.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Goodness of fit test based on modified likelihood ratio

Jutaporn Neamvonk, Bumrungsak Phuenaree
Department of Mathematics, Faculty of Science, Burapha University, Thailand

Testing whether a dataset followed a specified distribution is an important part of statistical analysis. There are many tests commonly used in the present, such as Chi-square, Kolmogorov-Smirnov, Anderson-Darling, Cramer von mises, Shipiro Wilk, and so on; however, none of these is most powerful for every set of data; even though, the Kolmogorov-Smirnov and Anderson-Darling test are usually found in many statistical software.

The test statistics, such as Kolmogorov-Smirnov, Anderson-Darling, Cramer von mises, were initially developed by weighted sum of a statistic, Z_t, which could be replaced by Chi-square statistic. Zhang (2002) proposed three new statistics, corresponding to Kolmogorov-Smirnov, Anderson-Darling, Cramer von mises, by replacing the Z_t with likelihood ratio statistic. The results showed that the powers of the new tests were superior to original tests.

In this research, we develop a new statistics test based on modified likelihood ratio statistic proposed by Cressie & Read (1984). It is applied to test whether the specified distributions are Gamma and Log-normal. The critical values of the test are investigated using randomly generated numbers in various distributions. The study of power compares the performance of Kolmogorov-Smirnov, Anderson-Darling, Zhang (2002)-Anderson-Darling and the proposed test. The result shows that the proposed test is superior than other tests when the generated random numbers are Logistic and Log-logistic distributions.

.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Towards Fail Safe Methods of Analysis, how do we balance the optimisation of marginal gains, with the potential large loses from rare events.

Tim Drye
Data Analysts User Group

Whilst fully trained and experienced statisticians are engendered with a healthy dose of skepticism about the outputs of their results. In particular we are fully aware, that conventional methods implicitly assume that unusual events are very rare indeed. Everything meaningful happens within 3 standard deviations.

However, particularly in circumstances involving human behaviour, typically were the assumptions that individual data points are independent as hopeful rather than likely, unusual events are much more common than conventional methods allow for. Typically this might be accounted for by applying a power law distribution rather than one that is exponentially based. However these methods become much more cumbersome and harder to interpret, they rarely get used outside specialist fields.

This paper advocates that we must encourage practitioners, to model those things in a more open way that makes more transparent the fragility of models that are overconfident about uncertainty. This is partly to counter the common characteristic of confirmation bias.

One strategy might be to begin by modelling rare but unwanted events, before common and desired benefits, keeping these as separate functions rather than merging into a single utility function. This could help detected when risks are growing within the parameter space even when benefits are still prevalent.

Another strategy might be to attempt to formulate an approach that turns the principles of maximum likelihood on its head, these approaches are inherently pessimistic about uncertainty, the principle is based upon seeking to minimise this uncertainty. However, approaches in other fields for example the enumeration of the explore-exploit dilemma and coupled behaviour with game theory show that flexibility and stochastic behaviour around a static optimum is benefitial. Can some principle of "minimum lossiness", accomodate a more optimistic approach to uncertainty and delivery more resilient outcomes.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Progression towards open access environments in the teaching of statistics to undergraduate medical students

Margaret MacDougall
University of Edinburgh

Introduction: The need for development of statistical competence within evolving medical curricula is a challenging one to address when faced with competing interests of relevance to preparing undergraduate medical students for safe clinical practise. However, if recommendations expressed by medical governance bodies to develop critical thinking skills are to be properly understood, this need must be recognized as universal across medical schools.

Methods and Results: The author has recently used WordPress as a tool for presenting and further developing her statistical knowledgebase resources specifically for undergraduate medical students engaged in short-term curricular research projects. The resultant site, StatsforMedics, is now available on an open access basis. Additionally, with funding support from the HEA and the University of Edinburgh Principal's eLearning Fund and Principal's Teaching Award Scheme and in collaboration with Learning Technologists, the author has developed statistics resources involving a range of clinical contexts. These resources are designed to make a contribution to steering undergraduate medical students through the conceptual maze of statistical concepts on the road to evaluating risk and improving their understanding of summary statistics and confidence intervals. A range of design features within these resources has been informed by the author's knowledge of the learning needs of students with Specific Learning Difficulties. She has also led development work involving the segregation of learning content into chapters, with the goal of making statistical learning more amenable to integration with clinical learning throughout undergraduate medical curricula. Through recent advances in technology at the University of Edinburgh, this output is now available in the form of open access resources.

Conclusions: Medical educationalists are invited to explore the new open access resources available to develop student competence in use of statistics. The author would welcome suggestions as to how she can enhance student learning experiences through further development of these resources.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

The importance of being honest

Steve Martin-Drury
Office for National Statistics

Many social surveys are well established, with methodologies and mechanisms in place for the production of statistics in a consistent manner, year in and year out. When a new survey year is begun in these 'production surveys', the path of least resistance (often followed) is to assume that, as the methodology, sample design and associated estimation processes were deemed appropriate in the previous year, therefore they should be in the current year. This poster seeks to explain the fallacy behind this supposition, and the ways in which established procedures should be challenged and alternatives explored regularly, even in well established production surveys. In this study, I will describe the work of my branch in looking at the methodologies behind various social surveys both for ONS and other Government departments, and how we embrace the need to regularly review and reassess the statistical theory behind each survey we administer.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Multivariate Statistics for Analysis of Honey Bee Propolis

Abdulaziz Alghamdi
Strathclyde University

Honey bees play a significant role ecologically and economically, through pollination of crops. Additionally, honey can be considered as one of the finest products of nature, with a wide range of beneficial uses, including use in cosmetictreatment,eye diseases, bronchial asthma and hiccups. Honey bees also produce beeswax, royal jelly and propolis.

Propolis is a resinous bee product, which consists of a combination of beeswax and resins which have been gathered by honey bees from the exudates of various surrounding plants. It is used by the bees to seal and maintain the hives, but is also an anti-infective substance which may protect them against disease. Propolis possesses a highly resinous, sticky gum appearance and its consistency changes depending on the temperature. It becomes elastic and sticky when warm, but hard and brittle when cold. Furthermore, its colours vary from yellowish-green to dark brown, depending on its age and source.

The purpose of this research is to use statistical analysis to study the biochemical properties of propolis, which have attracted much attention. Biochemical analysis of propolis leads to highly multivariate metabolomics data. The main benefit of metabolomics is to generate a spectrum, in which peaks correspond to different chemical components, making possible the detection of multiple substances simultaneously. Relevant spectral features may be used for pattern recognition.

This work will investigate the use of different statistical methods for analysis of metabolomics data from analysis of propolis samples using Mass Spectrometry (MS). Methods studied will include pre-processing methods and multivariate analysis techniques such as principal component analysis (PCA), clustering methods and partial least squares (PLS) methods. Background material and initial results of data analysis will be presented from samples of propolis from beehives in Scotland.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Modelling Climate Variables

Olaniyi Mathew Olayiwola, Oyeleke Ridwan Olaoye
Federal University of Agriculture, Abneokuta, Nigeria

Climate change is a change in the statistical distribution of weather patterns that lasts for an extended period of time. It is one of the most important issues facing humanity. Most scientists agree that emissions of anthropogenic greenhouse gases are responsible for the observed increases in global air temperature. Effects of both anthropogenic and natural causes of climatic change and their possible solutions were examined. The monthly weather data on solar radiation, maximum temperature, minimum temperature, rain, wind, dew point, 2-metre temperature and relative humidity were obtained from the Nigerian Meteorological Agency Weather Station. The eigenvalues and eigenvectors for the data set were obtained to determine the principal components. The bar and scree plots were used to select the required principal components. The correlation matrix was determined for the selected principal components and the atmospheric condition variables. The first principal component is strongly correlated with maximum temperature, minimum temperature, wind, dew point and relative humidity. The first principal component is primarily a measure of humidity as it correlates mostly with dew point and relative humidity with correlation values of 0.933 and 0.932 respectively. The second principal component is strongly correlated with solar radiation, maximum temperature, minimum temperature, wind, dew point and 2-metre temperature ,it is the measure of solar radiation as it correlates with it at 0.942. The third principal component correlates strongly with solar radiation, maximum temperature, wind and 2-metre temperature. As solar radiation decreases; maximum temperature, wind and 2-metre temperature increases. It is a measure of temperature as it has a correlation value of 0.909 with 2-metre temperature. This climatic pattern shows a steady increase in monthly average temperature which poses risks to human health and activities. High temperature is associated with risk of injury, illnesses, death from the resulting heat waves; wildfires, intense storms and floods rises.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

The use of a normal finite mixture model for the modelling of social and material deprivation in the Czech Republic

Ivana Malá
University of Economics, Prague

In the Survey of Health, Ageing and Retirement in Europe (SHARE) the European population aged over 50 is of interest; SHARE provides a multidisciplinary approach that delivers the comprehensive picture of the ageing process in Europe. In the fifth wave of the survey (carried out in 2013) two composite indicators of deprivation are given. The answers to the questions from the survey questionnaire are weighted, the index of material deprivation (11 items) and the index of social exclusion (15 items) on the scale from 0 to 1 are given. In this contribution, the empirical distributions of these indices in the Czech Republic in 2013 are analysed. Only weak dependence on the age is shown together with the stronger impact of education and (mainly) of the size of household. A mixture of two normal distributions (social exclusion index) and a mixture of one Bernoulli distribution (no deprivation) and two normal distributions for the index of material deprivation are fitted into the data. The hurdle model is applied to model the mixture of one discrete and two continuous distributions. Maximum likelihood method is used for the estimation of unknown parameters and all computations are performed in the program R.

Indices of material deprivation are regularly published by the Czech Statistical office and results for age groups 50-65 and above 65 are given. The indexes based on SHARE data are consistent with those given by the state statistics (according to our analysis).

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Structural Identification and Variable Selection in High-Dimensional Varying-Coefficient Models

Wing Kam Fung
Department of Statistics and Actuarial Science, The University of Hong Kong

Varying-coefficient models have been widely used to investigate the possible time-dependent effects of covariates when the response variable comes from normal distribution. Much progress has been made for inference and variable selection in the framework of such models. However, the identification of model structure, that is how to identify which covariates have time-varying effects and which have fixed effects, remains a challenging and unsolved problem especially when the dimension of covariates is much larger than the sample size. In this article, we consider the structural identification and variable selection problems in varying-coefficient models for high dimensional data. Using a modified basis expansion approach and group variable selection methods, we propose a unified procedure to simultaneously identify the model structure, select important variables and estimate the coefficient curves. The unique feature of the proposed approach is that we do not have to specify the model structure in advance, therefore, it is more realistic and appropriate for real data analysis. Asymptotic properties of the proposed estimators have been derived under regular conditions. Furthermore, we evaluate the finite sample performance of the proposed methods with Monte Carlo simulation studies and a real data analysis.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Text Analytics using Deep Learning Networks

Arindam Chaudhuri
Samsung R & D Institute Delhi

Deep Learning Neural Networks have direct correspondence to generalized linear models and basis regression functions in statistics. This is an insight that is useful in representing deep learning networks and an interpretation that does not depend on various hypothesis and processing based on which the human brain functions. The training process is based on maximum likelihood estimation encompassing large datasets functioning at very large-scale and real-world systems. A statistical perspective on deep learning caters and reflects to broad knowledge set that can be swapped between the two domains with the previlage of better understanding the regression problems.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

An evaluation of intermediate endpoints for gating of early phase clinical trials with some applications in cancer immunotherapy

Markus Elze, David Dejardin, Xian He, Hsin-Ju Hsieh, Daniel Sabanes Bove
F. Hoffmann-La Roche AG

There is renewed interest by sponsors and regulators to use intermediate endpoints (IME) for early phase clinical trials of drugs. A gating decision for further development frequently has to be made based on limited patient numbers and follow-up time, making the use of overall survival (OS) as primary study endpoint unfeasible in most cases. IMEs that mature early and are suitable for the drug's mechanism of action can be helpful to gate the next phases of development. This issue is particularly relevant for cancer immunotherapies, due to late separation of survival curves, challenges in using classic proxies such as progression-free survival (PFS) to predict ultimate OS benefit, and the need to consider the possibility of tumour pseudoprogression.

Data on responder status and initial lesion shrinkage are available early, while data on lesion change dynamics and progression take longer to mature. Combinations of several endpoints can also be considered, either as a composite endpoints or using dual endpoint gating.

In this study, we provide an overview of several established and novel IME, such as objective response rate, duration of response, time in response, durable response rate, and depth of response. We discuss their relative merits and shortcomings. By simulating early phase trials using data from larger late phase trials (including cancer immunotherapies), we investigate the association of IMEs with OS and the robustness of decisions made based on IMEs. We discuss maturity of the IMEs and availability of historical controls.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Box-Cox transformation for regression models with random effects

Amani Almohaimeed, Jochen Einbeck
Durham University

Regression analyses can help us detect trends, examine relationships and draw meaningful conclusions from experimental data. However, the assumptions of normality and of homoscedasticity of the response must be fulfilled prior to starting to analyze the data. The aim of this work is to ensure the validity of a normal distribution using the Box-Cox power transformation. The extension of the transformation to the linear mixed effects model was proposed by Gurka M. et al. (2006), in the case of a Gaussian random effect distribution. An obvious concern of assuming a normal random effect distribution is whether there are any harmful effects of misspecification. This problem can be avoided by estimating this distribution through the use of a technique known as "Nonparametric Maximum Likelihood", which, for our purposes, is adapted towards a "Nonparametric Profile Maximum Likelihood" technique. The feasibility of the approach is demonstrated through examples, including details of how to employ this approach in R.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Relative efficiency of randomized designs based on multiple comparisons

Abimibola Oladugba
University of Nigeria, Nsukka

This work is on the relative efficiency of randomized complete block design (RCBD) to completely randomized design (CRD) based on multiple comparisons. Two types of multiple comparisons, ad hoc and priori pairwise multiple comparisons were used to determine the relative efficiency using Scheffe’ type and Bonferroni simultaneous confidence intervals under the respective types of multiple comparisons. Both the ad hoc and the priori pairwise multiple comparisons results showed that the RCBD is more efficient than the CRD.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Bayesian Inference for Continuous Time Markov Chains

Randa Alharbi
University of Glasgow

In systems biology, understanding and analysing the system can be achieved by collaboration between the experimental work and mathematical modelling. Mathematical models can provide a better understanding about the system behaviour. However, modelling biological system can be challenging. The Mathematical model can be built based on existing knowledge and some assumptions. A well designed model requires to select suitable mechanism which can define the main component of the system and represent an appropriate law that can define the interactions between its components. The set of mechanisms that are used to describe the interaction between systems competent based on some unknown parameters which need to be inferred. The parameter can be estimated within the Bayesian framework. We defined stochastic repressilator system with Continuous Time Markov Chain (CTMC) model. CTMC is used as a stochastic model that describes the repressiltor system. Once the CTMC is constructed and the data are simulated, we aim to perform inference of model parameter. The difficulty when working with CTMC is that the evaluation of likelihood is intractable. Therefore, a key challenge in estimating parameters is choosing appropriate methods that allow us to perform the inference in case of intractable likelihood. Various approaches have been proposed to estimate model parameter within Bayesian context in such cases. We consider a synthetic biological system repressilator to perform inference. Two common proposed methods are selected to perform inference as an appropriate method for repressilator model. The first approach is approximate Bayesian computation (ABC) that allows to simulate from the model given the proposed value of parameters. The second approach is Particle Marginal Metropolis-Hastings (PMMH) which is one of particle MCMC family of algorithms that is based on sampling from target distribution. These sampling methods used to estimate model parameters. Finally, we compare the results we obtained from both methods.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

An alternative nonlinear growth models for biological processes

Oluwafemi Oyamakin, Angela Chukwu, Adebayo Bamiduro
University of Ibadan, Nigeria

Studies have shown that majority of the growth models emanated from the Malthusian Growth Equation (MGE), which is limited to growing without bounds. This study was designed to develop alternative growth models flexible to enhance internal prediction of biological processes based on hyperbolic sine function with bound. The intrinsic rate of increase in the MGE and its variants were modified by considering a growth equation, which produces flexible asymmetric curves through nonlinear ordinary differential equations. Weight of Japanese Quail; *Coturnix coturnix* L. (JQ) and Malaysian Oil Palm Fresh Fruit Bunches (MOPFFB), Top Height (NTH) from a Norwegian thinning experiment, sample plot 3661, *Gmelina arborea* Roxb. (GH), Pine (*Pinus caribaea* Morelet) (PH) and the diameter at breast height of Pine (*Pinus caribaea* Morelet) (PDBH) from organisations were used to test the validity of the new models in terms of general fitness and internal predictive status as well as robustness. Mean Square Error (MSE), Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and Residual Standard Error (RSE) were used to determine the best models among the proposed and existing models. The developed hyperbolic growth models captured boundedness in Malthusian Growth Equation, improved general fitness and robustness over exponential, monomolecular, Gompertz, Richards and von Bertallanffy growth models.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

An analysis of pilot whale vocalisation activity using Hidden Markov Models

Valentin Popov, Roland Langrock, Stacy DeRuiter, Fleur Visser
University of St Andrews

Vocalizations of cetaceans form a key component of their social interactions. Such vocalisation activity is driven by the behavioural states of the whales, which are not directly observable, so that latent-state models are natural candidates for modeling empirical data on vocalisations. In this paper, we use hidden Markov models to analyze calling activity of long-finned pilot whales (*Globicephala melas*) recorded over three years in the Vestfjord basin off Lofoten, Norway. Baseline models are used to motivate the use of three states, while more complex models are fit to study the influence of covariates on the state-switching dynamics. Our analysis demonstrates the potential usefulness of hidden Markov models in concisely yet accurately describing the stochastic patterns found in animal communication data, thereby providing a framework for drawing meaningful biological inference.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Decision analysis and uncertainty analysis for early test evaluation

Sara Graziadio, Kevin Wilson
Newcastle upon Tyne Hospitals NHS Foundation Trust

In early diagnostic test evaluations the potential benefits of the introduction of a new technology in the current healthcare system are assessed in the challenging situation of limited empirical data. These evaluations provide tools to evaluate which technologies should progress to the next stage of evaluation.

In this study a diagnostic test for patients suffering from Chronic Obstructive Pulmonary Disease (COPD) was evaluated with Bayesian networks, which provide a compact visualization of probabilistic dependencies and interdependencies. The structure of the network was inferred from the care pathway, a schematic representation of the journey of a patient in the healthcare system. After the network was inferred and reduced with arc reversal techniques, it was populated using expert judgement elicitation. The Bayesian network was then queried to evaluate whether the introduction of the test could reduce unnecessary hospital admissions. Uncertainty analyses were used to determine credible intervals for the comparison between the current and new pathway, and to identify influential parameters of the decision problem.

We found that the adoption of the diagnostic test had the potential to reduce the number of missed COPD exacerbations of symptoms that could lead to late hospital admissions, and of unnecessary visits to A&E. The model inputs that most influenced the posterior distribution were identified as the probability that a patient would go to A&E if an exacerbation was suspected, the probability that the healthcare professionals in primary care refer patients to the hospital, and the sensitivity of the test.

These results are useful to companies to inform the choice of the target population, of potential early adopters and the identification of the technological focus to guide development of the test.

The proposed approaches to care pathway analysis in early test evaluations could be a useful and intuitive tool for test evaluators, health economists and regulators.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Practical methods to pool multi-study joint longitudinal and time to event data

Maria Sudell, Ruwanthi Kolamunnage-Dona, Catrin Tudur Smith
University of Liverpool

Background: Joint longitudinal and time-to-event data models have been established in a single study case as beneficial compared to separate longitudinal or time-to-event analyses in a range of cases, including data with study dropout, time-to-event models with longitudinal covariates measured with error, or cases when the relationship between longitudinal and time-to-event outcomes is of interest. However the methodology available for multi-study cases such as meta-analyses is limited.

Aims: To investigate different approaches of modelling of multi-study joint longitudinal and time-to-event outcome data.

Methods: Several methods are examined to account for between study heterogeneity, including as one stage methods that can include random effects at the study level, stratification of baseline hazard by study and use of fixed study indicator terms and their interactions with treatment assignment, or approaches for two stage pooling of joint model fits. These methods are applied to a real data example and further investigated in a simulation study. Software have been developed in R to allow these methods to be easily applied in future investigations, which will be available in a package alongside joineR collaboration.

Results: The results from the real data example and simulation study will be presented at conference.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

A calibration method for non-positive definite covariance matrix in multivariate data analysis

Chao Huang, Jianxin Pan, Daniel Farewell
Cardiff University

Covariance matrices that fail to be positive definite arise often in covariance estimation. Approaches addressing this problem exist, but are not well supported theoretically. In this work, we propose a unified statistical and numerical matrix calibration, finding the optimal positive definite surrogate in the sense of Frobenius norm. Our proposed approach is implemented through a straightforward screening algorithm. Simulation results showed that the calibrated matrix is typically closer to the true covariance, while making only limited changes to the original covariance structure. We also revisited two substantive analyses to demonstrate the properties of the proposed calibration. This approach is not constrained by model assumptions. Neither is it limited by data structures. Since it is a calibration approach, it can be incorporated in existing covariance estimation process, and offers a routine check and calibration of covariance matrix estimators.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Variance Estimation Methods in Meta-Analysis: An Appraisal

Mona Pathak, Sada Nand Dwivedi, Bhaskar Thakur
All India Institute of Medical Sciences, New Delhi

Meta-analysis is a statistical technique for synthesizing the effect size of considered outcomes from various comparative studies. Heterogeneous group of studies are often encountered, which need to be synthesized by random effect method. The results under random effect method rely on the weights involving calculated between study variance. There are various methods to estimate the between study variance (τ^2).The present study was aimed to assess between study variance using various methods and its appraisal.

There are various methods of between study variance estimation: (i) Method of moments; (ii) Cochrane's ANOVA method(CA); (iii) DirSimonian and Liard Method(DL); (iv) Paule & Mandel estimate; (v) DerSimonian and Kacker methods using CA and (vi) DerSimonian and Kacker methods using DL. These methods were compared using systematic review of randomized controlled trials comparing efficacy of neoadjuvant chemotherapy in comparison to adjuvant chemotherapy among breast cancer patients. A total of 17 eligible studies, out of 1239 identified records through PubMed and Cochrane register of controlled trials were used for application considering overall survival(OS), recurrence free survival(RFS), loco-regional recurrence(LRR) and distal recurrence(DR). For comparison, pooled effect estimate through random effect method using each of above-mentioned variance estimation methods was obtained. For further appraisal, bias & mean square error (MSE) of point estimate; and coverage probability of corresponding interval estimates were used.

The considered outcomes, OS, RFS, LRR and DR were reported by 14, 10, 10 and 11 studies; with varying heterogeneity (I^2) of 0%, 39.3%, 14.3% and 46.7% respectively. With changing variance estimation methods, the estimate of τ^2 also varied providing different weight and corresponding varied pooled effect estimate. The comparative results will be presented and discussed in detail. Random effect method is sensitive to calculated between study variance. Hence, its method of calculation needs to be considered in view of pros and cons of comparative results.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Use of online change-point detection methods for space weather storms

Qingying Shu, Marian Scott, Lyndsay Fletcher, Matteo Ceriotti Ceriotti, Peter Craigmile
University of Glasgow

Space weather describes the environmental conditions in space determined by plasma physics and solar processes. Magnetic storms are one of the most important space weather phenomena, and defined as large-scale electromagnetic disturbances in the near-Earth space environment. Magnetic storms may damage satellites and sensitive technology systems, so early detection is an important step in mitigating their effects.

We first simulate the near-Earth magnetic field under storm and quiet conditions using a physically-based semi-empirical model, with existing satellite trajectory data as time and position inputs. We then apply online prospective detection algorithms to the simulated spatio-temporal data with the goal of identifying the storm onset as accurately and quickly as possible. Storms may change the mean, variance or both of the magnetic signal. Two criteria for comparing the detection methods are the average run length and the mean delay. Our results show that for detecting change in scale, the Mood non-parametric test outperforms the Bartlett test under a Gaussian assumption.

Using the physical positional constraints on Earth-orbiting satellites, and the characterisation of detectability of storms, we propose network design considerations for near-earth magnetic storm detection.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Investigating honey bee colony losses from surveys of beekeepers

Alison Gray, Magnus Peterson
University of Strathclyde

Honey bees are both economically and ecologically vital, owing to their key role in pollination of agricultural crops and hence food security, improving both quality and quantity of yield, as well as pollination and propagation of many other plants. A healthy honey bee colony is also a source of honey, pollen, beeswax, propolis and royal jelly.

Honey bees currently face many threats, including lack of forage diversity owing to intensified agriculture, pests and diseases, effects of agricultural pesticides, and severe weather conditions. Sudden widespread large-scale colony losses, occurring mostly in the USA in winter 2006/7, provoked a huge amount of research worldwide to study the level of colony losses of honey bees and investigate potential risk factors for colony loss.

In Scotland we have been collecting questionnaire data on colony losses over winter since 2006, via surveys of beekeepers, and are now in our 10th year of these surveys run from the University of Strathclyde. These use geographically stratified random sampling and a purpose-designed questionnaire. Since 2010 we have contributed data to the colony loss monitoring research of COLOSS (Prevention of honeybee COlony LOSSes; www.coloss.org), an international research association studying honey bee colony losses since 2008. Data analysis involves descriptive statistics, generalised linear mixed modelling, multivariate and spatial analysis.

Some results on beekeeping and colony losses will be presented both for Scotland and internationally.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Beekeeping and colony losses in Saudi Arabia

Abdulmajeed Albarrak, Alison Gray
Department of Mathematics and Statistics, University of Strathclyde, Glasgow

The most commonly managed honey bee is *Apis mellifera*, a highly adaptable species found almost worldwide and a highly valuable pollinator of agricultural crops. The total number of managed honey bee colonies worldwide was reported to have increased by 64% from 1961 to 2007. However Europe, North America, and some other parts of the world have since 2006/7 observed sharp declines of managed honey bee colonies. A series of surveys reported that many colonies suddenly disappeared, leaving no or only a few remaining living bees, referred to as “Colony Depopulation Syndrome” (CDS) or “Colony Collapse Disorder” (CCD). Various reported possible causes include parasitic mites, malnutrition, harsh winters, exposure to neonicotinoid pesticides, or a combination of these. No single factor has gained widespread acceptance.

The research network COLOSS (Prevention of honeybee COlony LOSSes) was formed in 2008, to investigate honey bee colony losses, and carries out annual loss monitoring surveys in many countries in Europe and beyond. However, CCD has not yet been reported in the Arabian Peninsula and colony loss data are limited. Honey bee races in the Arabian Peninsula are morphologically dissimilar to the European honey bee, use of agricultural pesticides there is low, and infestation by *Varroa* and other pests is less prevalent. Therefore, investigation of colony losses in the Arabian Peninsula will add to the body of research into causes of colony losses.

This work is amassing information on the nature and infrastructure of beekeeping in Saudi Arabia, with the aim of establishing a new survey of beekeepers there to study beekeeping management and experience of colony losses. Results so far will be reported.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Gamma- generalized gamma mixture cure fraction model in survival data

Serifat Folorunso
Department of Statistics, University of Ibadan

In this study, we use a methodology based on the Gamma Generated link function in the presence of mixture cure fraction models, considering that survival data are skewed in nature. The objective of the study is to propose a skewed family distribution on mixture cure model using a skewed Gamma link function which can handle heavily skewed survival data. The mathematical properties of this proposed model were explored and the inferences for the models are obtained. The proposed model called Gamma- generalized gamma mixture cure model (GGGMCM) will be validated in the presence of a real life survival data set from Gynecology Oncology Unit of University College Hospital, Ibadan, Nigeria.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Misconception of HIV/Aids Among Ever-Married Women and Associated Factors: A Comparison of Non- Spatial Multilevel and Spatial Model

Majida Jawad, Sohail Chand, Jawad Kadir
University of the Punjab

Combating HIV/Aids is getting importance in the world health concerns as it increases with the passage of time. Pakistan is also under the attack of the HIV/ Aids virus therefore combating HIV is also the big health concern in Pakistan. Combating HIV is only possible by giving sufficient awareness to the general public especially women of child bearing age i.e. 15-49 years and the core hindrance of combating HIV is the misconception about HIV transmission. Therefore, combating HIV and misconception are associated to each other. Therefore, the present study aimed to identify the spatially distribution of the three type of misconception factors of HIV (i.e. transmitted by mosquito bite, supernatural means and sharing food with HIV positive person) and also to determine the core factors that may affect in reducing the misconception and increases the chances of combating HIV in Pakistan. Multilevel, Bayesian Multilevel and Hierarchical Spatial Autoregressive models were applied to the data and results from them revealed that the Hierarchical Spatial Autoregressive models is more appropriate. The results revealed that the condom use, women age, education of household head, hepatitis were the significant factors for all three types of misconception about HIV.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Application of Spatio-Temporal Regression Model with Partial Differential Equations Regularization approach on Brain Data

Salihah Alghamdi
University of Glasgow

Recently, much attention has been paid to Brian data that study the brain activity and provide better understanding of the brain. Electroencephalography (EEG) is a non-invasive technique that used for monitoring and recording the electrical activity of the brain. EEG data consists of repeated measurements observed at different locations of the brain over some time interval. We take a functional data analysis approach to analyse the EEG data. In particular, we extend Spatio-Temporal Regression Model with Partial Differential Equations Regularization approach and provide a classification framework for repeated measurements EEG data. We apply our methodology on training and test data, which record an individual EEG signals while viewing two different kinds of images as a stimulus.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Bayesian Spatio-Temporal Analysis of Small Area Typhoid in Zhejiang Province, China

Xiuyang Li, Zihang Li, Mengyin Wu, E Marian Scott
University of Glasgow

Objectives Explore Bayesian methods to analysis the spatio-temporal evolution of typhoid fever risk pattern in Zhejiang Province, China from 2005 to 2015.

Methods The research represents the first application of Bayesian spatio-temporal modeling with spatial auto-correlation and temporally auto-regression to analysis typhoid fever. The Bayesian model, fitted by Markov Chain Monte Carlo simulation using WinBUGS, stabilized risk estimates in small areas and controlled for spatial auto-correlation.It estimated (1) area-specific risk pattern; (2) year-specific risk pattern; and (3) Posterior probabilities of area-specific risk ratio (RR) differing from away from the average level for revealing locations of hot/cold spots.

Results Typhoid fever exhibited a declining RR trend across the study region during the past eleven years.Variation of area-specific RR was statistically significant, which was apparent from the map of (95% credible interval) of RR. Hot spots in the north east and south east, cold spots in west of the region were identified, and all areas are cold spot in 2014 and 2015.

Conclusions Bayesian spatio-temporal analysis contributes to a detailed understanding of small-area typhoid trends and risks. It estimates RR for each area as well as an overall average trend.The approach of identifying hot/cold spots through analyzing and mapping probabilities of area-specific typhoid fever trends and RR differing from the mean trend high-lights specific locations where typhoid fever situation is deteriorating or improving over time. Future research should analyze trends more regions and/or more periods (allowing for non-linear time trend) and risk factors.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

The perfect recipe for continued success – a case study in ensuring a sustainable supply of shellfish across the globe

Sophie Carr, Peter Keen
Bays Consulting

The impact of illegal, unregulated and unreported fishing is having a devastating impact on fish stocks across the globe. Many people depend upon the oceans for food and the ability to sustainably manage food resources s is of critical importance to the wellbeing of communities around the world.

Conserving and managing fish and shellfish stocks is therefore a high priority. Establishing the origin of landed shellfish is critical for the enforcement of no-take zones, assuring regional branding of high value and creating sustainably maintained stocks. However, when a catch landed can law enforcement agencies actually prove where the shellfish were taken from?

Over the last few years, Keen Marine and Bays Consulting have been part of a study into the use of carbonate microchemistry in mussel shells to determine if mussels could be assigned to the specific location in which they were grown. All the mussels in the study were sourced as spat from a single origin outside of the study area, and were cultured in three harbours with similar underlying geochemistry, but different surrounding land use.   When harvested and the shells analysed for sixteen different trace elements.

The analysis was undertaken without prior knowledge of where the three locations were, or which sample had been grown in which location to prevent analyst bias influencing the analysis. Initial results indicate not only improvements in the methodology for the chemical analysis of the shells, but also determined a microchemical “fingerprint” unique to mussels grown at each of the different locations. The results of this study have now being developed into a larger feasibility study to support evidence based policy making for protecting coastal fisheries and developing long term sustainable food sources.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Initial insights into service data from 'Closer to Home,' community initiatives to prevent frail, elderly hospital admissions in NHS Forth Valley

Maria Cristina Martin, Matt-Mouley Bouamrane, Kimberly Kavanagh, Paul Woolman
University of Strathclyde

NHS Forth Valley's 'Closer to Home' (C2H) programme covers a range of community services aimed at preventing hospital admissions for frail and elderly people. These include Enhanced Community Teams (ECTs) which integrate health and social care staff to support patients at home, and the Advice Line for You (ALFY), a nurse-led telephone service providing health and social care advice or simply reassurance, both implemented in December 2015. Here, the main objective is to provide initial insights into the first year of the ECTs service.

The majority of community service data within NHS Forth Valley is collected through their Multi-Disciplinary Information System (MiDIS). Data linkage techniques using Structured Query Language (SQL) have been used to retrieve and link together data sources pertaining to the ECTs held within NHS Forth Valley's MiDIS database.

During the first year of the service (Jan-Dec 2016), 544 patients were referred on to the ECTs. 89% of these patients had only one episode of care onto the service, while the majority of the rest were admitted to the service a second time (604 episodes in total). Only one patient was admitted a third time during the year. Of referrals where information was supplied on referrer, 59.7% of referrals were made by a GP and 15.6% were made by discharge coordinators at hospital. Other sources included nursing services, Allied Health Professionals and ALFY. 30% of referrals had the primary referral reason of facilitating a discharge from hospital by supplying a package of care. This was not originally an intended purpose of the service. Other referral reasons will be presented.

These initial insights into the C2H service have proved useful in informing the design of data capture of the service and will form an essential baseline for future work evaluating the impact of this initiative in a robust framework.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Career planning for statistics students as a part of a course in statistics

Kadri Meister
Umeå University

Students applying for a bachelor’s degree program in Statistics at Umeå University, Sweden, have usually heard about the possibility of working (almost) anywhere as a statistician, but lack a clear idea what a career in statistics is all about. The aim of the course "Statistics Application areas” is to introduce a large variety of statistical application areas to the students. During the course, students have the opportunity to visit potential working places in Umeå and in Stockholm. During these visits, the students meet statisticians from different application areas, and can learn about the need for statistical knowledge (eg. if there is a need to know any specific statistical methods) and their every-day work (eg. do you work mostly independently or in a group?). While planning these visits, we teachers put great effort on finding a good mixture of different working places; including administrative authorities as well as companies in the private sector.

In addition to the above, the course also includes exercises where the students critically examine results and conclusions in different statistical work, and discuss ethical issues related to these. The course also includes exercises how to communicate statistical concepts, issues and results with non-statistician. In the end of the course, every student chooses an application area for deeper studies and writes a short paper about a statistician's work in the chosen field of application.

Students say that they appreciate the course because they get to see the width of different working areas for statisticians. They also claim that the course provides them with new insights into possible career paths and motivates them to more actively design their further education based on their own interests. In summary, “If you do not know how to work as a statistician, then that course is very good." (From the course evaluation 2017).

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Frances Wood: first female medical statistican

Tim Cole
UCL Great Ormond Street Institute of Child Health

This year the Society awards the Wood medal for the first time. It commemorates the distinguished but little known statistician Mrs Frances Wood (née Chick) who died in 1919 aged 35. Her life was by any standards remarkable – born in 1883 to a family of seven girls who all had distinguished careers, at a time when professional opportunities for women were limited. In 1908 she graduated in chemistry from UCL and started research at the Lister Institute of Preventive Medicine. The arrival in 1910 of Dr Major Greenwood to set up the Lister Statistics Laboratory caused her to switch to medical statistics, from which point her career rise was meteoric.

Between 1913 and 1916 Frances published six papers, alone (including a JRSS read paper) and with Greenwood, on the statistics of index numbers and cause-specific mortality rates. During the First World War she worked at the Board of Trade, latterly the Ministry of Munitions, where she was awarded the MBE, then OBE, “for services in connection with the war”. She became a Fellow of the Royal Statistical Society in 1913, was elected to RSS Council in 1915 and to the Executive Committee in 1917–the first woman to hold either post.

Frances married Sydney Wood in 1911, and in 1919 gave birth to their daughter Barbara; however she died just 12 days later of birth complications.

Soon after her death the Society invited Fellows and others to donate to the Frances Wood Memorial Prize. The call raised £300 (£14,000 now) and the prize was first awarded in 1921 “for the best investigation of any problem dealing with the economic or social conditions of the wage-earning classes.” It lapsed in the 1980s but has been reinstated as the Wood medal, a timely reminder of a remarkable woman.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Using the FMT estimator for analysis of censored household demand data

Maria Karlsson, Thomas Laitila
Umeå University

The FMT estimator is earlier suggested for estimation of censored regression models (see Karlsson and Laitila, 2014). It is based on maximum likelihood estimation of a finite mixture of Tobit models, and defined as a weighted sum of component regression parameters. Here, simulation results on properties of three covariance matrix estimators are presented. Also, a finite mixture of heteroskedastic Tobit models is suggested for analysis of consumer demand with data characterized by censoring. The resulting estimator is evaluated in a simulation study and an empirical application to household expenditure data.

References:

Karlsson, M., Laitila, T. (2014), Finite mixture modelling of censored regression models, *Statistical Papers* **55**, 627-642.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Weighted samples in random forests

Amirah Alharthi, Charles Taylor, Jochen Voss
University of Leeds

Classification trees and random forests methods are two powerful predictive models that are used for extracting knowledge from data. The random forest has the following stochastic components: the variable selection and sampling within or without replacement in the sample size. In this research we consider the input of using weighted samples, as a way to control, or increase randomness in each tree in the random forest. We aim to investigating different components with a view to a better understanding about what makes good performance.

Poster & Awards Reception
Tuesday 5 September – 6.15pm-7.30pm

Rugby player performance and injury risk management

Fang Guan, Lesley Walls, Matthew Revie
University of Strathclyde Business School

Rugby is a sport which involves heavy contact between players resulting in frequent injuries, impairing individual player and team performance. Researchers have explored effective methods to predict injuries (Gabbett, 2016; Kiesel, Butler and Plisky, 2014; Rowson and Duma, 2013; Myer et al., 2010). Weaknesses of these studies include ignoring the dynamic nature in the fixed training load threshold of players and a lack of consideration of subject effects in modeling process. Our research aims to address this, by developing a dynamic forecasting model which takes account into player’s subjective effects based on a less complete data set than that has been used previously. The research is a joint project between the University of Strathclyde and the governing body of rugby union in Scotland, the Scottish Rugby Union (SRU). The data used in this study comes from three data sets: a player injury data set, which records player's injury history in the latest seasons, a player training load data set, which records player daily training details and a player wellness monitoring flat data set, which records player's daily wellness conditions. In total, 16,753 observations have been analysed. Although a rich number of observations are available for this study, for many of these observations, data are incomplete.

This research will start from a systematic review of methods of data mining from incomplete and especially multivariate data, followed by a review of modeling methods on subjective effects of participants. Further, one phenomenon we identified is the hysteresis nature in player's response and injury occurrence to training loads. And thus methods on how to capture hysteresis in the forecasting model and form a fit forecasting period is another area we will explore. Research gaps on these methods will be identified by our study.

4.1 Contributed – Medical Statistics: Pragmatic methods and design
Wednesday 6 September – 8.45am-9.45am

Analysing Stepped Wedge Cluster Randomised Controlled Trials by means of a Linear Mixed Effects Model: A simulation study

Alecia Nickless, Ly-Mee Yu, Merryn Voysey
Nuffield Department of Primary Care Health Sciences, University of Oxford

A stepped wedge cluster randomised trial (SWCRT) is a special case of a cross-over randomised controlled trial, where the cross-over only occurs from the control condition to the intervention condition, and clusters are randomised to the time of cross-over. A simulation study was undertaken to compare different formulations of a linear mixed effects model to analyse data from a SWCRT under different scenarios of temporal change in the outcome and different intervention effects. The purpose of this analysis was to determine the best formulation of the linear mixed effects model to correctly estimate the intervention effect in the presence of temporal change in the outcome. Multiple datasets were simulated under 22 different scenarios, including scenarios with no trend, linear trend and non-linear trend in the outcome over time. Nine different formulations of the linear mixed effects model with either compound-symmetrical or autoregressive covariance structures were considered as potential approaches to analysing these data.

Including continuous terms for time and for time exposed to the intervention most frequently resulted in the correct conclusion for the intervention effect. Models with multiple indicator variables used to represent intervention and time lapse at each time step, such as when categorical terms for both time step and exposure time are included in the model, lead to poor estimates of the parameters. Models with no interaction term between time and the intervention did not account for intervention-specific effects on the outcome trend over time, thus the predicted outcome could not be derived from the model at a specific point in time, after a known exposure time to the intervention.

When analysing SWCRTs we recommend that linear mixed effects models include time as a continuous variable, and include terms for the intervention and either for the interaction between intervention and time, or a term for exposure time.

4.1 Contributed – Medical Statistics: Pragmatic methods and design
Wednesday 6 September – 8.45am-9.45am

Bivariate network meta-analysis for surrogate endpoint evaluation

Sylwia Bujkiewicz, Dan Jackson, John Thompson, Rebecca Turner, Keith Abrams, Ian R White
University of Leicester

In early regulatory decision-making, evidence on effectiveness of new health technologies may be limited due to the long follow-up time required to measure their effect on final clinical outcome. To expedite such decisions, shorter-term surrogate endpoints can be used if they are good predictors of clinical benefit across multiple trials in different populations and/or of different treatments. Candidate surrogate endpoints, however, are often not perfect, and the level of association between the treatment effects on surrogate and final outcomes may vary between treatments.

Bivariate random effects meta-analysis (BRMA), which synthesises jointly correlated effects, can be used to predict the treatment effect on the final outcome from the treatment effect measured on a surrogate endpoint. However, BRMA does not differentiate between the treatments, as it assumes a common distribution for the true treatment effects across all treatment contrasts. Network meta-analysis (NMA) combines data from trials investigating heterogeneous treatment contrasts and has the advantage of estimating effects for all contrasts individually by assuming exchangeability of the true treatment effects across studies within the same treatment contrast only. We exploit this framework to model the association between treatment effects on surrogate endpoints by the use of bivariate NMA. Modelling assumptions about the between-studies heterogeneity and the network consistency and their impact on predictions are investigated using simulated data and examples in multiple sclerosis and colorectal cancer.

When the association between the treatment effects on surrogate and final outcome is weak across all studies of all treatments but strong within particular treatment contrasts, bivariate NMA can perform better than BRMA in predicting the treatment effect on the final outcome from the effect on the surrogate endpoint in a new study. When data are sparse, additional exchangeability assumptions may be necessary in order to make predictions, for example about the effect of a new treatment.

4.1 Contributed – Medical Statistics: Pragmatic methods and design
Wednesday 6 September – 8.45am-9.45am

Using published Kaplan-Meier curves to reconstruct survival data for secondary analyses

Yinghui Wei, Patrick Royston
Plymouth University

The hazard ratio is often recommended as an appropriate effect measure in the analysis of survival outcome from randomised controlled trials. In meta-analysis of aggregated survival data across trials, an essential step is to extract the (log) hazard ratio and its variance from published trial reports. However, in some trial reports, hazard ratios are not reported but the Kaplan-Meier curves are available. In this work, we develop a tool to reconstruct the necessary survival data to enable the approximation of hazard ratios. The reconstructed survival data also open the possibility to assess the proportional hazards (PH) assumption and estimate alternative effect measures if the PH assumption is violated. We will illustrate the application of this reconstruction tool to several examples, and assess the accuracy of the approximation by comparing the summary statistics between the reconstructed data and the original publications.

4.2 Contributed - Official statistics and public policy: Harnessing Routine Data
Wednesday 6 September – 8.45am-9.45am

Using routine administrative records to comprehensively assess the non-fatal and fatal burden of 132 conditions in Scotland

Grant Wyper, Ian Grant, Oscar Mesalles-Naranjo, Elaine Tod, Colin Fischbacher, Gerry McCartney, Diane Stockton
NHS National Services Scotland

Introduction

Previous efforts to quantify the comprehensive burden of disease in the Scotland have relied on modelled data from other countries. This aim of this study is to utilise routine data sources to provide a transparent and systematic approach to assess the population health loss due to 132 conditions.

Methods

Records on each death in Scotland were used alongside life expectancy data to calculate the Years of Life Lost to premature mortality (YLL) as a measure of fatal burden. To quantify the non-fatal burden, we utilised data across a wide range of healthcare services such as consultations with GPs, community prescriptions and hospital attendances in the inpatient, psychiatric, outpatient and unscheduled care settings. The record linkage of datasets through a common patient identifier allowed for a thorough search of clinical contacts to provide refined estimates. A range of coding systems were used to define each condition (READ, ICD, BNF, OPCS). Combining these estimates with the Global Burden of Disease 2015 study's relative assessment of disability for each condition, we calculated the Years Lived with Disability (YLD). Combining YLL and YLD provides an estimate of the number of Disability-Adjusted Life Years (DALY), which encapsulates the full burden of non-fatal and fatal disease.

Results

Approximately 50% of health loss in 2015 was due to fatal causes and was higher for males. Conditions responsible for generating the highest fatal burden were Ischaemic Heart Disease, Lung Cancer and Chronic Obstructive Pulmonary Disease. The non-fatal burden was largely due to mental health and conditions that were associated with old age, poor diet, obesity, tobacco, alcohol or illicit drug use.

Conclusion

These findings are important to help policy makers and planners to identify the drivers behind health loss. Understanding geographic, demographic and socioeconomic variations provides evidence for the stratified management of healthcare resources and interventions.

4.2 Contributed - Official statistics and public policy: Harnessing Routine Data
Wednesday 6 September – 8.45am-9.45am

300 years Wargentin birth: TABELLVERKET's data analysis

Elisabeth Morand, Nathalie Le Bouteillec
Ined

In 1749 the Swedish kingdom instigated the *Tabellverket* (translated into the Standing Committee on Tables) with the purpose of gathering population statistics, i.e. to measure the population size, its main characteristics etc. This was the first formal statistics body to be formed in Europe.

The scientists of the Royal Academy of Sciences (including Carl von Linné) pushed for this ambitious project and Pehr Wargentin, an astronomer and secretary of the Academy, contributed to the interpretation of the results. He wrote several memoires on population statistics. Those memoires are not known but they are essential in the history of statistics. As this year, it is 300 years anniversary of Wargentin's birth, we would like to take this opportunity to underline the fabulous work done by this scholar focusing on the survival function and from there Wargentin had a vision of the age pyramid.

As mentioned above Pehr Wargentin focused in particular on the survival function. Contrary to the current utility of the age pyramid, which aims to establish average durations of survival, the survival function was a way for the scholars of the time to deduce the total population. To approximate the universal multiplier or to find a more exact method to estimate the size of the population they constructed mortality with the parameters they knew as the number of births or the number of deaths. The calculations made by Wargentin and previously by Hayley are based on a stable and closed population. In this situation, the calculations allow to represent a survival curve and also a pyramid of age if one inverses x-axis and the y-axis.

The aim of this communication is to present the pyramid according to Wargentin and others views of the time as well as to put it into perspectives with actual graphical representations.

4.2 Contributed - Official statistics and public policy: Harnessing Routine Data
Wednesday 6 September – 8.45am-9.45am

Occupied address (household) estimates from Administrative Data

Pete Jones
Office for National Statistics

The Office for National Statistics is looking beyond the next census in 2021 to the possibility of an Administrative Data Census based on linking together administrative data held by the Government.

Our main goal is to produce as many census-type statistics as possible by combining administrative and survey data, and comparing them with the outputs from the 2021 Census. To demonstrate our progress in reaching this goal, we have established a series of 'Research Outputs' which also demonstrate new opportunities from such an approach.

Our first attempt at producing household estimates focused on a concept based on 'occupied addresses' from administrative data. This paper will outline the methodology we have used to derive these estimates by linking address records from administrative sources to Unique Property Reference Numbers (UPRNs).

We highlight the challenges of using administrative data, with consideration for the following:

- the quality of automated matching of address information held on administrative datasets
- difficulties in aligning administrative data with traditional Census definitions of 'household'
- the impact of lags on administrative data following changes of address
- limited availability of information about the relationships between individuals resident at the same address
- the coverage and classification of address types, including communal establishments and homes in multiple occupation

Key to the development of household statistics from administrative data is establishing suitable methods to combine with survey data. We will outline the methodological framework for combining administrative data with surveys to adjust for biases in our estimates for the number of households, with some early results. We will also show our progress towards producing modelled estimates of household size and composition for local authorities.

We conclude with a discussion about our need to understand user requirements for future household statistics, and the potential to produce new outputs with administrative data.

4.3 Contributed - Environmental/Spatial statistics: Climate Applications
Wednesday 6 September – 8.45am-9.45am

Reducing uncertainty in low-frequency meteorological hazards for the nuclear industry

Paul Newell, Nicolas Fournier, Simon Brown
Met Office

The nuclear industry typically needs to consider the resilience of its infrastructure to low-frequency hazards of the order of 1-in-10,000 annual probability of occurrence. Extreme value models provide the framework for estimating such rare events, allowing extrapolation to quantiles well outside the observed record. In some situations, these extrapolations can lead to physically unrealistic values, reducing confidence in results, especially when combined with large levels of uncertainty at extreme levels. Such results also drive a significant increase in engineering costs for new-build nuclear assets that will need to factor in these values as part of the design criteria and health and safety case. The added uncertainty of climate change poses additional challenges to designers.

One approach to address these challenges is to explore available climate model data to increase the amount of available information and thus reduce uncertainty. The latest version of the Met Office high-resolution climate model provides a `virtual` event set consisting of over 1400 years (40 ensembles of 35 years) of daily scenarios, many times larger than the available observed record. This means that it contains many more physically plausible extreme events so that the extent of extrapolation, and thus uncertainty, may be greatly reduced.

This approach has already been successfully applied for extreme rainfall for the UK National Flood Resilience Review (NFRR; https://www.gov.uk/government/publications/national-flood-resilience-review).

In addition, the multivariate extremes model of Heffernan & Tawn (2004) allows the definition of joint extreme events that may be easier to design for than consideration of single hazards independently.

The event set and associated analysis methods are thus ready to answer a very wide range of questions related to the risk landscape for design and operation of a proposed nuclear plant.

4.3 Contributed - Environmental/Spatial statistics: Climate Applications
Wednesday 6 September – 8.45am-9.45am

Nonparametric and semiparametric decomposition of time series and space-time processes, with applications to climate data

Jan Beran, Britta Steffens
Department of Mathematics and Statistics, University of Konstanz, Germany

Certain long series of space-time observations show a changing seasonal pattern. Occasionally, the observed changes may coincide with known events. Often however, the reasons are less obvious. In case of climate for instance, changing patterns can have practical consequences on various natural processes as well as, health, economics, and others, typically related to adaptation, vulnerability and various risk factors (see reports of the IPCC). We decompose our space-time processes into trend, a smoothly changing seasonal component and stationary residuals. A remarkable result then emerges. The asymptotic convergence rate of the estimated seasonal component is unaffected by the strength of dependence in the data. This happens even if there is long-range dependence. This is in sharp contrast to nonparametric estimation of the trend. The asymptotic variance of the seasonal component depends however on the unknown spectral density at harmonic frequencies. In this talk, we discuss some background theory and describe a new bandwidth selection algorithm and model fitting methods for the residuals. The proposed methods are applied to some climate data, including wind speed, temperature and precipitation. Not only the trend but also changes in the seasonal variability over time, in particular the amplitudes, turn out to have a strong spatial component.

4.3 Contributed - Environmental/Spatial statistics: Climate Applications
Wednesday 6 September – 8.45am-9.45am

Modelling short time series of annual abundance indices as a function of high-dimensional weather data

David Elston, Mark Brewer, Pete Henrys
Biomathematics and Statistics Scotland

The need to improve our understanding of species-weather relationships is of increasing importance in order to identify species which are likely to come under particular pressure from climate change. However, whilst many studies have indicated associations between weather and population size, our knowledge of these associations is restricted by the statistical tools used.

For any particular species and any chosen weather variable (e.g. rainfall, temperature), we have developed an approach to modelling annual abundance indices in which the mean for each abundance index is a linear combination of covariates comprising monthly weather records from the 12 months of the index year and many more monthly covariates from preceding years. Including all covariates in a single model enables a detailed exploration of weather-abundance relationships, but at the cost of having more regression coefficients than observations of the response variable.

By specifying the regression coefficients for successive months as a non-linear function of a small number of underlying parameters, we can overcome the limitations of having many more covariates (months) than abundance indices (years). The chosen function constrains the sequence of regression coefficients for successive months to follow a damped Fourier oscillation. This construction combines periodicity in the influence of seasonal variation in weather on species abundance with a reducing influence of weather on abundance at increasing time lags.

This talk will describe the specification of the models, discuss the challenges of automating the model-fitting process to enable application to a large number of species, and indicate recent refinements aimed at increasing the precision of estimation by allowing for known but uncontrolled variation in the observation process underlying each annual abundance index.

4.4 Contributed - Social statistics: Record linkage and big data
Wednesday 6 September – 8.45am-9.45am

Measuring Big Data skills: Two proxy measures

Alana McGuire
University of Stirling

Big Data has become somewhat of a buzzword in the past couple of years, with research showing an apparent increase in demand for Big Data skills (E-skills UK, 2014). However, many questions still remain around Big Data: How can we measure Big Data skills? Is Big Data creating new skill divides? Is region important in examining data skill shortages? Are certain sectors more likely to experience data skill shortages? Does the size of the organisation impact data skill shortages? This talk aims to explore and answer these questions.

This paper presents findings from my PhD research and proposes two proxy measures for Big Data skills. The measures use data from the Employer Skills Survey, the Labour Force Survey, and the British Cohort Study. The first proxy is in the form of a Big Data ‘base skills score’ variable, composed from several skill shortage variables in the Employer Skills Survey which are asserted to be important in the literature for managing and analysing data. Data from the Labour Force survey is merged in for analysis to provide sector averages for gender division. The second proxy uses arithmetic data from a follow up to the 1970 British Cohort Study and merges this with later socioeconomic data from the 2004/05 follow up study to examine gender, ethnic, and social inequalities in Big Data skills.

In this paper, the composition of both variables will be explained and some multilevel regression models will be presented using the two proxy measures. I will conclude that both measures are plausible, and useful proxies for Big Data skills.

References

E-Skills UK. (2014) *Big Data Analytics: Assessment of Demand and Labour and Skills 2013-2020.*Accessed online at https://www.thetechpartnership.com/globalassets/pdfs/research-2014/bigdata_report_nov14.pdf

4.4 Contributed - Social statistics: Record linkage and big data
Wednesday 6 September – 8.45am-9.45am

Combining limited data on income tax with household surveys: A simulation-based approach for estimating income inequality in India

Sunil Kumar
King's College London

This paper proposes an approach to augment survey data on Indian household incomes with limited statistics on income tax released by the government in order to obtain better estimates of income inequality. I use a simulation-based procedure to explore the implications of missing information about the distribution of incomes in both data sources. This procedure yields bounds on the resulting estimates of income inequality which can be clearly interpreted in terms of the assumptions used to combine the two sources of data. The results suggest that levels of income inequality are higher than those estimated from either data source on its own.

4.4 Contributed - Social statistics: Record linkage and big data
Wednesday 6 September – 8.45am-9.45am

Innovative use of census data to study variation in the health of ethnic groups: the Scottish Health and Ethnicity Linkage Study

Duncan Buchanan, Raj Bhopal, Geneviève Cezard, Esta Clark, Anne Douglas, Laurence Gruer, Andrew Millard
NHS National Services Scotland

Like the rest of the UK, Scotland is becoming more ethnically diverse and there is a duty on public bodies to promote equality across ethnic groups. This has implications for health service policy and provision but reliable data on service use and outcomes for ethnic groups are difficult to achieve relying on administrative health data alone. In Scotland, an innovative data linkage research programme has been instrumental in addressing this gap in recent years.

The Scottish Health and Ethnicity Linkage Study (SHELS) set out, with appropriate information governance, to link the 2001 census data to NHS patient registration data to allow analysis of routinely collected health records based on self-reported ethnicity from the census. This has provided a unique retrospective cohort of 4.62 million people, representing 91% of the resident population, which has been used to analyse a range of health outcomes. The most recent phase of the study, based on 12 years follow up, included comparison of all-cause hospitalisation rates and mortality rates using Poisson regression, adjusting for age, socio-economic status and country of birth. The first direct estimates of life expectancy by self-reported ethnicity in the UK were generated using the revised Chiang method.

The results showed a generally healthier population among minority ethnic groups in comparison with the White Scottish majority, in part reflecting a possible 'healthy migrant' effect. However as in earlier phases of research, notable variations in outcome and service use were evident across ethnic groups, independent of socio-economic status. SHELS has shown that secure analysis of confidential census records and administrative health data is practical and achievable for wider public benefit. The results have important implications for policy, planning and clinical care to tackle health inequalities in Scotland and beyond.

4.5 Contributed - Methods and theory: **Trials & Surveys**
Wednesday 6 September – 8.45am-9.45am

Calibrating Non-Probability Samples with Probability Samples Using LASSO

Chen Jack, Michael Elliott, Richard Valliant
Survey Monkey

One of the most prominent applications of survey research is election polling. The timeframe to collect critical voting intention is short, typically spanning just the last few weeks prior to the election day. Due to declining land-line phone coverage and improved phone screening technology, it has become a significant challenge for election pollsters to capture voting intentions in a timely way. This has led to the expanded use of non-probability samples, particularly cheap and easily accessed samples of individuals obtained from web surveys. But non-probability samples are at risk for selection bias due to differential access, degrees of interest, and other factors. Calibration is a standard method used in official statistics and other settings that uses weights to adjust total estimates from a sample to known totals in a population. Because non-probability samples do not have robust inferential properties, we consider use of model-assisted calibration methods that allow robust estimation of population totals. In particular, we consider calibration to estimated population totals using adaptive LASSO regression – estimated-controlled LASSO (ECLASSO). Adaptive LASSO can yield a consistent estimator of a population total as long as a subset of the true predictors is included in the prediction model, thus allowing large numbers of possible covariates to be included without risk of overfitting. This allows to the possibility of calibration to estimates from higher-quality probability samples with modest sample sizes. We apply ECLASSO to predict the voting spread (proportion of Democratic votes minus the proportion of Republican votes) for 11 gubernatorial elections and 8 senate elections in the U.S. 2014 midterm election. Since the actual election results are published, we can compare the bias and root-mean square error of ECLASSO with traditional weighting adjustment methods.

4.5 Contributed - Methods and theory: **Trials & Surveys**
Wednesday 6 September – 8.45am-9.45am

Bayesian Inference for Population Attributable Risk

Sarah Pirikahu, Geoff Jones, Martin Hazelton, Cord Heuer
Massey University

Epidemiologists often wish to determine the population impact of an intervention to remove or reduce a risk factor. Population attributable type measures, such as the population attributable risk (PAR) and population attributable fraction (PAF), provide a means of assessing this impact, in a way that is accessible for a non-statistical audience. To apply these concepts to epidemiological data, the calculation of estimates and confidence intervals for these measures should take into account the study design (cross-sectional, case-control, survey) and any sources of uncertainty (such as measurement error in exposure to the risk factor).

We provide methods to produce estimates and Bayesian credible intervals for the PAR and PAF from common epidemiological study types and assess the Frequentist properties. The model is then extended by incorporating uncertainty due to the use of imperfect diagnostic tests for disease or exposure. The resulting model can be non-identifiable, causing convergence problems for common MCMC samplers, such as Gibbs and Metropolis-Hastings. An alternative importance sampling method performs much better for these non-identifiable models and can be used to explore the limiting posterior distribution.

The data used to estimate these population attributable measures may include multiple risk factors in addition to the one being considered for removal. Uncertainty regarding the distribution of these risk factors in the population affects the inference for PAR and PAF. To allow for this we propose a methodology involving the Bayesian bootstrap. We also extend the analysis to allow for complex survey designs with unequal weights, stratification and clustering.

4.5 Contributed - Methods and theory: **Trials & Surveys**
Wednesday 6 September – 8.45am-9.45am

The Efficiency and Optimality characterization of certain balanced incomplete-block designs emanating from some quasi-semi-Latin squares

Polycarp Chigbu, Eugene Ukaegbu
University of Nigeria, Nsukka, NIGERIA

Balanced incomplete-block (BIB) designs were constructed from three types of three-factor block-structured combinatorial designs which are not necessarily semi-Latin squares but whose symbols (treatments) are *ab initio* arranged as in the semi-Latin square, presented by Chigbu (1999), and which were called quasi-semi-Latin squares (QSLS) in Chigbu (2009, 2012), are given. The basic consideration for the construction involves the exploitation of the different block structures of the experimental units of the quasi-semi-Latin squares for the same treatments (symbols). The statistical properties of the resulting equireplicate binary designs were evaluated. Hence, the pairwise and canonical efficiency factors, and some related optimality criteria concepts were utilized in the designs' evaluations. The percentage loss of information when estimating corresponding basic (simple) contrasts of all the possible block structures due to each type of QSLS is also calculated. All the emanating designs are found to be connected but only the designs in nine blocks are established to be both variance-and efficiency-balanced. Also, with the estimation of their respective basic contrasts, all the designs in nine blocks offer gain in efficiency over those of unblocked designs of equivalent sizes but lead to about 46 percent loss of information when compared with all the designs in six blocks but one.

4.6 Contributed - Communicating and teaching statistics: Communicating and presenting statistical results
Wednesday 6 September – 8.45am-9.45am

Choosing the Right Angle: A Conscientious yet Pragmatic Approach to Avoiding a Fishing Expedition

Neil Spencer
University of Hertfordshire

John Tukey said Statisticians “get to play in everyone's back yard” and for an applied statistician, the variety of topics with which one may become involved and the expertise and enthusiasm of those with whom one works are some of the best aspects of the job.

At the same time, statisticians working in a research team or acting as a consultant may be the only statistical experts involved with a project. They are faced with the difficulty of balancing the desire for scientific rigour with the aspiration of the research team or client to gain as much as possible from the data in terms of “exciting” insights and/or publications.

Whilst researchers who have had some statistical training in research methods courses are aware that trawling the data for “statistically significant” results is not appropriate, they also recognise that there may be interesting discoveries to be made even after the primary analysis, for which the data were collected, has taken place. However, for the statistician involved, an exploration can easily turn into what they would regard as a “fishing expedition”. The researcher/client finds interesting patterns and wants to carry out an assessment of how likely these are to be real or due to chance fluctuations in the data. They undertake hypothesis tests and if these do not yield small p-values, attention naturally turns to other potentially interesting patterns.

This talk describes how the author, faced with the above situation, has developed a pragmatic approach to the issues, simultaneously satisfying his research collaborators’ desires for an exploration of a dataset for interesting patterns and his own need to ensure that analyses retain statistical rigour. Issues and solutions discussed involve the generation of hypotheses, significance thresholds and sigma levels, publication bias and appropriate ways of reporting the results of such analyses.

4.6 Contributed - Communicating and teaching statistics: Communicating and presenting statistical results
Wednesday 6 September – 8.45am-9.45am

When do confidence intervals and p-values give different interpretations of statistical significance? An investigation into 2x2 tables

Nick Beckley, Fiona Reid
King's College London

Analysing a 2x2 table is one of the longest-studied problems in statistics, and one of the first problems encountered by statistics students. For large samples, approximate confidence intervals (CIs) and p-values usually give matching interpretation of statistical significance (that is, a (1-α) CI contains the null value if and only if p>α). For small samples, exact p-values are often reported without a corresponding exact CI. This is in part due to many statistical programs not reporting matching exact CIs with exact p-values as standard, combined with a general emphasis on reporting the p-value over a CI for 2x2 tables in research and statistics education. However in many research fields we often want to report an effect size, for which a CI is more appropriate, and there has been a recent drive to emphasise the reporting of CIs over p-values in medical and psychological journals. Therefore it is vital that methods for calculating matching CIs for exact p-values are widely available to applied statisticians.

We investigate discrepancies between exact p-values and commonly used CIs for risk differences (RDs), risk ratios (RRs) and odds ratios (ORs) within small samples. We perform simulations of all 2x2 tables with at least one ‘small’ cell for sample sizes of up to 100 per group. The distribution of p-values from these simulations are presented separately for significant and non-significant CI interpretations. We find substantial discrepancies with exact p-values when using common CIs for RDs, as well as some discrepancies for common CIs for RRs and ORs.

These results are of particular interest for proponents of reporting absolute risk differences in medicine, as researchers may be put off reporting contradicting p-values and CIs. We suggest possible solutions to these issues, including highlighting some less traditional, but still well-established, methods that could be promoted for mainstream statistical usage.

4.8 Contributed - Official statistics and public policy: Brexit and foreign investment
Wednesday 6 September – 8.45am-9.45am

UK Trade Statistics - Post-EU referendum Measurement and Analysis

Adrian Chesson
Office for National Statistics

This presentation explains the main components and outputs covered within the Office for National Statistics' UK Trade statistics, why these are so important for understanding our economy in a global context and how the demands and priorities for these statistics have changed since the UK's vote to leave the EU on 23 June 2016. The presentation will focus on the priorities following the referendum and published in the ONS UK Trade development plan. Particular attention will be given to work to enhance our use of new source data to provide more granular statistics and also our collaboration with academic partners to better understand UK Trade asymmetries.

The need for more granular data and analysis will initially be met through better use of existing data and by increasing the sample size for the International Trade in Services (ITIS) survey. Also, improving the ITIS sample design to optimize it on geography will provide better-quality information for individual countries. We are seeking to identify new sources of price data and make more use of administrative data sources to produce better-quality deflators. The presentation will expand on these developments.

Trade asymmetries exist across global trade statistics whereby, for example, the exports recorded by country A to country B do not match the imports recorded by country B from country A. There are multiple reasons such as measurement differences, conceptual differences and different data sources. We have conducted some analysis of the UK's trade asymmetries, engaged in bilateral meetings with other National Statistical Institutes (NSIs) to understand some of the potential reasons for differences and are currently working with an academic to develop possible models that might lead to improvements. This work is on-going, but we will share results where possible.

4.8 Contributed - Official statistics and public policy: Brexit and foreign investment Wednesday 6 September – 8.45am-9.45am

Bubbles, blind-spots and Brexit

John Fry, Andrew Brint
Sheffield Hallam University

In this paper we develop a pre-existing financial model to investigate whether bubbles were present in opinion polls and betting markets prior to the UK's vote on EU membership on June 23rd 2016. The importance of our contribution is threefold. Firstly, our continuous-time model allows for irregularly spaced time series -- a common feature of polling data. Secondly, we build on qualitative comparisons that are often made between market cycles and voting patterns. Thirdly, our approach is analytically tractable. Thus, where bubbles are found we suggest a suitable adjustment. We find evidence of bubbles in polling data. This suggests they systematically over-estimate the proportion voting for remain. In contrast, bookmakers' odds appear to show none of this bubble-like over-confidence. However, implied probabilities from bookmakers' odds appear remarkably unresponsive to polling data that nonetheless indicates a close-fought vote.

4.8 Contributed - Official statistics and public policy: Brexit and foreign investment Wednesday 6 September – 8.45am-9.45am

Ultimate controlling parent of businesses (UCP) analysis and what it tells us about the ultimate origin of foreign direct investment (FDI)

Yanitsa Petkova, Sami Hamroush, Michael Hardie
Office for National Statistics

The Office for National Statistics (ONS) currently publishes foreign direct investment (FDI) statistics on immediate basis in line with international guidance. According to this measure in 2015 the European Union (EU) accounted for 45% of the FDI positions held by foreign companies in the UK. Half of these EU positions were held by companies in the Netherlands and Luxembourg which are renowned financial centers - where investment is generally in transit and intended for an ultimate destination. Estimating how much of this investment is genuinely coming from these two countries and where the rest of it is ultimately originating from is essential for understanding the importance of the EU to the UK economy and for informing the Brexit negotiations. There is also a wide user interest in the UK foreign direct investment statistics produced by their ultimate point of origin from both UK and international institutions.

The ONS is currently working on producing FDI statistics by their ultimate origin making the UK one of only a few countries in the world which is developing this kind of challenging and innovative statistics. The investment estimates are produced by linking ultimate parent data held on the Interdepartmental Business Register (IDBR) to inward FDI (investment in the UK). The IDBR holds information only on companies with over 50% ownership, while the FDI survey reports on companies with ownership over 10% which leads to some gaps in the UCP coverage. These gaps were filled by using the investment destination reported by the FDI's survey respondents. The findings of this work have not been published yet with the presentation at the RSS conference being one of the first places where the ultimate investment origin of FDI flows, positions and earnings will be discussed.

4.9 Invited - Prize Winners: Young Statisticians Meeting 2017
Wednesday 6 September – 8.45am-9.45am

Do nursing stations within bays of hospital wards reduce the rate of inpatient falls? – An interrupted time series analysis

Usama Ali, Andrew Judge, Angela Brooke, Lauren Davis, Katie James, Sallie Lamb
University of Oxford

Objectives: Falls are a major public health burden for individuals and society. Falls are associated with loss of independence, functional decline and are a contributing reason for subsequent admission to long-term care. The financial burden associated with injurious falls is substantial. This study aims to evaluate whether the introduction of portable nursing stations within bays of hospital wards, has led to a change in trend of the monthly rate of inpatient falls.

Methods: Data on inpatient falls from local hospital records (Datix) were collected monthly between April 2014 and March 2017 across 17 wards within two UK hospitals (Stoke Mandeville & Wycombe General). The outcome was the monthly rate of falls per 1000 occupied bed days (OBDs). Using a natural experimental study design, interrupted time series analysis was used to assess whether trends in fall rates changed following the introduction of portable nursing stations in April 2016.

Results: A total of 2322 falls were identified in the study period, with 59% being males. The overall median age was 81. The majority of falls (99.3%) were classified as either none, low or moderate harm with 0.5% resulting in severe harm and less than 0.2% resulting in death. Up to April 2016, the monthly rate of falls incfreased by 0.127 (p=0.002). After the intervention was introduced, the monthly rate of falls was decreasing by 0.437 (p=0.003). At 12 months post-intervention, the absolute difference in the rate of falls between the estimated post-intervention trend and the pre-intervention projected estimate was 4.59 falls per 1000 OBDs. This was a relative percentage reduction of 46.3%.

Conclusion: This study provides evidence that the introduction of SITB was temporally associated with a reduction in the monthly rate of falls. This has the potential to be applied across the NHS and may reduce the number of inpatient falls.

4.9 Invited - Prize Winners: Young Statisticians Meeting 2017
Wednesday 6 September – 8.45am-9.45am

Asymmetries in UK Trade Flows

Katie O’Farrell
Office for National Statistics

International trade data has long been characterised by asymmetries between importers and exporters measurements. For the UK, trade statistics are about to take centre-stage in decision making - analysis on and understanding of these asymmetries would allow trends in UK trade to be as informative as possible in advance of these decisions. Today's talk will discuss the potential methodological and economic causes of trade asymmetries; as well as present descriptive statistics on the UK's largest bilateral asymmetries for trade in goods. A key question raised by trade asymmetries is whether data reported by partner countries could be used as a check for trade data collected within the UK. Using a statistical approach outlined by Baranga (2017), a piece of work produced in collaboration with the UK Trade team, a model of estimating reconciliation weights, based on average variance in partner countries' reports, will be presented.

4.9 Invited - Prize Winners: Young Statisticians Meeting 2017
Wednesday 6 September – 8.45am-9.45am

Analysis of Network Time Series

Kathryn Leeming, Guy Nason, Marina Knight, Matthew Nunes
University of Bristol

A network time series consists of data collected over time at nodes of a network, or graph. These networks arise in a wide range of settings, such as environmental, social, and medical.

In this talk analysis of a network time series is presented using the NARIMA (Network ARIMA) framework. As an extension to the univariate ARIMA time series model, NARIMA allows for dependence on neighbouring nodes according to a (possibly changing) network structure.

Model fit and relevant statistics will be assessed to compare different models for our example data. Potential methods of handling common time series issues arising on networks will also be discussed, such as removal of temporal and spatial trend structure.

5.1 Invited - Economic evaluation in relation to infectious diseases
Wednesday 6 September – 9.55am-11.15am

Benchmarks of value in economic analysis: the use of cost-effectiveness thresholds

Paul Revill
University of York

Health care systems face considerable population health care needs with often severe resource constraints. The way in which available resources are allocated across competing priorities is crucial in affecting how much health is generated overall, who receives health care interventions and who goes without. Cost-effectiveness analysis (CEA) can assist policy-makers in resource allocation. The central concern of CEA is whether the health gains offered by an intervention are large enough relative to its costs to warrant adoption. This requires some notion of the value that must be realized by an intervention, which is most frequently represented using a cost-effectiveness threshold (CET).

CETs should be based on estimates of the forgone benefit associated with alternative priorities that consequentially cannot be implemented as a result of the commitment of resources to an alternative. For most health care systems these opportunity costs fall predominantly on health as a result of fixed budgets or constraints on health systems' abilities to increase expenditures. However, many CEAs to inform decisions have used aspirational expressions of value, such as the World Health Organization's (WHO) recommended CETs (of 1-3 times GDP per capita in a country) which are not based upon this kind of assessment. Consequentially, they do not reflect the realities of resource constraints and their use is likely to reduce overall population health and exacerbate health care inequalities.

This talk will illustrate these points and present recent and ongoing research to inform the choice of CETs. The research has implications for assessment of intervention cost-effectiveness, pricing and investments in implementation and delivery of health care.

5.2 Invited - A view from Scottish Statistics
Wednesday 6 September – 9.55am-11.15am

A view from Scottish Statistics

Esta Clark
National Records of Scotland

Scotland’s Census 2021 will be predominantly online which has implications for how we design, collect, process and disseminate the data. Esta Clark from the National Records of Scotland will talk through plans for the census in 2021 and also showcase some of the award winning infographic and data visualisation work to disseminate NRS Demographic Statistics.

5.2 Invited - A view from Scottish Statistics
Wednesday 6 September – 9.55am-11.15am

A view from Scottish Statistics

Maighread Simpson
NHS National Services Scotland

What does the future hold for the publication of health statistics in Scotland? Why not come along and find out! This session will explain how NHS Scotland's Information Services Division (ISD) are using user design, data science and agile development to develop a publication model which provides users with the data they need, in the way that they need it.

5.2 Invited - A view from Scottish Statistics
Wednesday 6 September – 9.55am-11.15am

A view from Scottish Statistics

Gregor Boyd
Scottish Government

Statistics.gov.scot takes a world leading approach to publishing official statistics as open data. Gregor Boyd from the Scottish Government’s Office of the Chief Statistician provides an overview of what the system can do.

5.3 Invited - Developments of functional data analysis for environmental sensor data Wednesday 6 September – 9.55am-11.15am

Nonparametric statistical downscaling for data fusion of in-lake and remotely-sensed chlorophyll-a data

Craig Wilkie, Marian Scott, Claire Miller, Andrew Tyler, Peter Hunter, Evangelos Spyrakos
University of Glasgow

Earth observation of lake water quality is becoming increasingly common but has a number of statistical challenges, including the calibration of the satellite images. Chlorophyll-a is a green pigment that provides an indirect indicator of lake health and can be quantified from satellite imagery, providing better spatial and temporal coverage than traditional in-lake sampling. However, these grid-scale data must be calibrated with in-lake data (assumed accurate within measurement error), which are sampled from the lake directly at a specified number of point locations and at distinct time points. In this talk, I will present a novel statistical downscaling model, which addresses the change-of-support problem in matching data at different spatial and temporal scales. This approach enables data fusion of the in-lake and remotely-sensed data, combining spatial and temporal information from the remotely-sensed data with accuracy from the in-lake data.

This novel approach, which involves using smooth functions over time within the statistical downscaling framework, enables a functional data approach which relates the resulting basis coefficients through a spatially-varying coefficient regression. A hierarchical Bayesian modelling approach additionally leads to comprehensive uncertainty quantification.

Data were provided by the GloboLakes project (www.globolakes.ac.uk), a consortium research project investigating the state of lakes and their responses to environmental change on a global scale. The model will be applied to an example dataset for Lake Balaton, Hungary, a large and shallow lake that has suffered from poor water quality in the past and is therefore of interest to water quality researchers. Data are available for 9 in-lake locations, sampled approximately fortnightly, and 7616 grid cells of remotely-sensed data, for 115 consecutive monthly averages. The utility of the model for successfully fusing log(chlorophyll-a) data of different spatiotemporal support will be demonstrated.

5.3 Invited - Developments of functional data analysis for environmental sensor data
Wednesday 6 September – 9.55am-11.15am

Spatio-temporal modelling of sparse remote-sensing data

Mengyi Gong, Claire Miller, Marian Scott
University of Glasgow

Remote-sensing technology is widely used in environmental monitoring, providing measurements with exceptional spatial coverage and resolution. For example, the ARC-Lake project (http://www.geos.ed.ac.uk/arclake) has processed retrievals from the Advanced Along-Track Scanning Radiometer instrument on ESA's Envisat platform to produce spatio-temporal lake surface water temperature (LSWT) data for lakes globally. However, such data can contain sparse images, at particular time points, due to cloud cover or instrument failure, presenting challenges to the statistical analysis. For such sparse remotely-sensed images, missing data imputation and dimension reduction are usually required as part of the spatio-temporal modelling. To overcome these challenges, we present a bivariate functional principal component analysis (FPCA) for the remote-sensing images, which is formulated as a mixed effect model, with a time-varying mean function (SS-FPCA).

FPCA is frequently used to reduce data dimensionality and identify variational patterns. In the presence of missing observations, a mixed model FPCA (MM-FPCA) can be used to overcome the computational problems due to missingness in the standard FPCA algorithm.

Our method introduces a time-varying mean function constructed using a state space model and obtained by the Kalman filter/smoother to additionally account for temporal dependence between remote sensing images. The SS-FPCA is estimated using the Alternating Expectation-Conditional Maximization algorithm. A simulation study shows that the model estimates are robust throughout the repetitions.

The SS-FPCA is then applied to the LSWT data for Lake Victoria, in central Africa, recorded on a 47 by 57 grid over 202 months. A tensor product of B-spline functions are proposed for the construction of the bivariate functional data and the spatio-temporal patterns are investigated using the SS-FPCA. The model reconstructions provide smaller RSS as compared to the MM-FPCA using the same degrees of freedom, especially in pixels where data availabilities are low.

5.3 Invited - Developments of functional data analysis for environmental sensor data Wednesday 6 September – 9.55am-11.15am

A clustering method for spatially dependent functional data

Elvira Romano
Department of Mathematics and Physics, Universit´a della Campania "Luigi Vanvitelli", Caserta, Italy

A common aim of monitoring georeferenced phenomena is to search for zones that have similar behaviours. In these cases the problem is to define clustering methods such that spatial dependence is taken into account. Inspite of several research initiative such as, hierarchical approaches, dynamic strategies, model based methods, and a purely spatial approaches, none of these solve an interesting and typical problem of environmental fields like the description of clusters in terms of spatial dispersion. We propose Dynamic Clustering (DC) method which classifies georeferenced curves according to their contribution to the spatial variability. In particular, we define a spatial dispersion function associated to each curve and propose a k-means like clustering algorithm. The performance of the proposed method is illustrated by an application on a real dataset about ozone concentration data in Eastern United States.

Keywords: clustering; spatially dependent functional data, ; spatial dispersion function.

References

[1] Delicado, P., Giraldo, R., Comas, C. and Mateu, J.: Statistics for spatial functional data: some recent contributions. Environmetric, 21: pp.224-239, (2010)

[2] Giraldo, R., Delicado, P., Comas, C., Mateu, J.: Hierarchical clustering of spatially correlated functional data. Statistica Neerlandica, (2011)

[3] Haggarty, R., Miller, C., Scott, E.M.: Spatially Weighted Functional Clustering of River Network Data. Journal of the Royal Statistical Society, Series C.,(2015)

5.4 Invited - Sustainable Development Goals
Wednesday 6 September – 9.55am-11.15am

Sustainable Development Goals

Glenn Everett
Office for National Statistics

Kim Bradford Smith
DFID

The Sustainable Development Goals (SDGs) are a global system of targets and indicators covering everything from Poverty to Infrastructure to Peace and Justice. For the first time they will cover developed countries like the UK as well as developing countries, and also look below national estimates at subgroups to ensure we as “leave no one behind” in achieving this ambitious agenda.

This session will include: the story of the political process that developed this hugely ambitious agenda; the ONS’ role in collecting, analysing, presenting and disseminating data for the UK; DfID’s work with the UN system and least developed countries to support this global ambition and some the novel technical challenges of measuring some of the most ambitious targets.

5.5 Invited - Papers from the Journal of the Royal Statistical Society: Networks
Wednesday 6 September – 9.55am-11.15am

Estimating whole brain dynamics using spectral clustering

Yi Yu, Ivor Cribben
University of Bristol

The estimation of time-varying networks for functional Magnetic Resonance Imaging (fMRI) data sets is of increasing importance and interest. In this work, we formulate the problem in a high-dimensional time series framework and introduce a data-driven method, namely Network Change Points Detection (NCPD), which detects change points in the network structure of a multivariate time series, with each component of the time series represented by a node in the network. NCPD is applied to various simulated data and a resting-state fMRI data set. This new methodology also allows us to identify common functional states within and across subjects. Finally, NCPD promises to offer a deep insight into the large-scale characterisations and dynamics of the brain.

5.5 Invited - Papers from the Journal of the Royal Statistical Society: Networks
Wednesday 6 September – 9.55am-11.15am

Respondent-driven sampling bias induced by community structure in social networks

Luis Rocha, Anna Thorson, Renaud Lambiotte, Fredrik Liljeros
Karolinska Institutet, Sweden

Some populations such as sex-workers and injecting drug users are difficult to sample because of the absence of a sampling frame. Respondent-driven sampling (RDS) has been proposed and extensively used as an alternative methodology to overcome this limitation. Similarly to snowball sampling, RDS uses peer-recruitment to reach hidden individuals. Since people with more acquaintances are sampled more often, RDS takes into account the number of social contacts of sampled individuals to adjust their contribution on the estimation of population level variables. RDS is a peer-referral method and therefore is constrained by the structure of the social network of the target population. Social networks are known to be non-random and in particular often display community structure. Network community structure refers to groups of individuals more connected between themselves than with individuals in other groups and may emerge due to various factors such as geographic, age group, or temporal constrains. In this presentation, we will discuss some biases induced by community structure on the so-called RDS-II prevalence estimator widely used in the literature. We model and perform a series of numerical simulations of RDS on both synthetic and empirical populations represented by their social networks aiming to reproduce real-life scenarios. We find that the estimated prevalence of a hypothetical variable is associated with the size of the network community to which the variable-positive individual belongs and observe that individuals with few contacts may be under-sampled if the sample and the network are of similar size. We also find that the RDS-II estimator performs well if response rates are relatively large and the community structure is weak, whereas low response rates typically generate strong biases irrespectively of the network community structure.

5.5 Invited - Papers from the Journal of the Royal Statistical Society: Networks
Wednesday 6 September – 9.55am-11.15am

Statistical clustering of temporal networks through a dynamic stochastic block model

Catherine Matias, Vincent Miele
CNRS - University Paris 6

Statistical node clustering in discrete time dynamic networks is an emerging field that raises many challenges. Here, we explore statistical properties and frequentist inference in a model that combines a stochastic block model for its static part with independent Markov chains for the evolution of the nodes groups through time. We model binary data as well as weighted dynamic random graphs (with discrete or continuous edges values). Our approach, motivated by the importance of controlling for label switching issues across the different time steps, focuses on detecting groups characterized by a stable within-group connectivity behaviour. We study identifiability of the model parameters and propose an inference procedure based on a variational expectation–maximization algorithm as well as a model selection criterion to select the number of groups. We carefully discuss our initialization strategy which plays an important role in the method and we compare our procedure with existing procedures on synthetic data sets. We also illustrate our approach on dynamic contact networks: one of encounters between high school students and two others on animal interactions. An implementation of the method is available as an R package called dynsbm.

5.7 Invited - Data Science for Public Good
Wednesday 6 September – 9.55am-11.15am

Sensing human behaviour with online data

Suzy Moat, Tobias Preis
Data Science Lab, Warwick Business School

Our everyday usage of the Internet leaves volumes of data in its wake. Can we use this data to help us reduce delays and costs in measuring human behaviour, or even to measure behaviour we couldn’t measure before? Here, we will outline a number of studies carried out at the Data Science Lab at Warwick Business School, investigating whether online data can help us monitor disease levels, measure global travel patterns, and evaluate whether the beauty of the environment we live in might affect our health. We will discuss some of the challenges in generating estimates of human behaviour from online data when our relationship with Internet services continues to evolve at such rapid pace.

5.7 Invited - Data Science for Public Good
Wednesday 6 September – 9.55am-11.15am

Data Science for public good - how ONS is building capability & providing new insights through data science; addressing statistical quality and ethical issues.

Peter Fullerton, Rowena Bailey
ONS

The goal of ONS's new Data Science Campus is to help build data science capability for the benefit of the UK, using a new generation of tools and technologies that exploit the growth and availability of innovative data sources. By applying data science skills to some of government’s and society’s big problems, we seek to extract the maximum research and public interest value from public data holdings, promoting analysis which is more relevant, timely and fit-for-purpose for the modern, global, connected world we live in and helping build the skills and capacity to manipulate and interrogate data sources of all types effectively.

This talk will describe the rapid journey to establish a high-performing data science team since the initial start-up in July 2016. The presentation will provide examples of the Campus’s work including applications of machine learning, image processing, economic analysis and public health analysis. The examples will illustrate some of the quality considerations and ethical issues being addressed.

As well as giving an overview of the broad span of Campus research projects, the talk will outline the innovative approach taken to enhance data science capability and the partnerships forged with government, research institutes, academic, commercial and international bodies.

5.7 Invited - Data Science for Public Good
Wednesday 6 September – 9.55am-11.15am

The Urban Big Data Centre - Delivering a Research-Led National Data Service

Andrew McHugh
Urban Big Data Centre, University of Glasgow

The Urban Big Data Centre was established in 2014 as part of ESRC's Big Data Network to provide research and data service functions to advance knowledge about urban settings and to inspire innovation. The Centre offers support, infrastructure, tools, training and knowledge exchange needed to use data to drive impact. Our ultimate aspiration is to make positive transformations in urban life, achieved by unlocking the power of data to improve policy and practice for social, economic and environmental well-being in cities.

This talk will describe several aspects of the Centre's work, covering its data collection building, research project development, data science and data management innovation and its networks of strategic partnerships with government, commercial, academic and international bodies. The talk will cover a range of legal, technical, procedural and organisational challenges, providing insights into how they are being managed at UBDC.

5.8 Invited - Exploiting Data to Manage Complex Assets
Wednesday 6 September – 9.55am-11.15am

Statistical simulation and modelling using Design of experiments and Data for New Product Development and Asset Management

Misti Paul
Advanced Analytics

A critical component of effective asset management is identifying and designing the right asset in the first place to best support the customer or end user's needs and support the organisation's growth. Good asset providers will always be looking to add or improve the features of their assets to best meet demand requirements and design new products to meet new requirements. Organizations are faced with tremendous challenges in designing new products. The products have to be designed optimally to meet the needs of the customer as well as from an asset growth perspective be a valuable addition to the portfolio of existing assets ensuring revenue growth and profit maximization.

In order to achieve this organizations have to determine the features to be included in a product and most importantly the optimal combination of features. The product cannot be seen and analysed in its entirety as whole, but decomposed to partial contributions (partworths) of product features. The method developed to do this is through conjoint analysis, using design of experiments to design an orthogonal design to test the optimal combination of product features. In this way, researchers can not only explain the preferences of existing products and assets but also simulate preferences for entirely new products that are defined by feature combinations. This paper will examine conjoint analysis in detail and how it's used in new product development with statistical simulation with some real life case studies.

5.8 Invited - Exploiting Data to Manage Complex Assets
Wednesday 6 September – 9.55am-11.15am

Exploiting Data to find out where you are: A case study in applying Bayesian statistics in the marine environment

Sophie Carr
Bays Consulting

In all walks of life, knowing where you are, and frequently more importantly where you are in relation to both where your assets are and should be is not always simple. Increasing easy access to geo referenced data can help - navigation aids are almost standard on smart phones whilst continual developments in analytics and computation means driverless cars become an ever closer reality. However, in many cases autonomy on land and in the air relies on the features such as roads to assist with navigation or a GPS signal. But what happens when there neither is available? It is this situation in which many autonomous marine based vehicles operate. This presentation provides an overview of a system designed with minimal battery and computational power requirements capable of providing autonomous marine vehicles with a position using only information extracted from images collected on commercially available cameras. Determining location and direction from multiple sparse, intermittent time based optical data sets provides an interesting case study of the application of Bayesian statistics within industrial research and trials.

5.8 Invited - Exploiting Data to Manage Complex Assets
Wednesday 6 September – 9.55am-11.15am

Simulating the Statistical Performance and Perturbation Effects on a Complex Asset Fleet Maintenance Plan

Nira Chamberlain
Babcock

The planned maintenance of a fleet of complex assets has a major impact on the cost and performance of any service provider, especially those with limited high value assets. In designing an asset schedule maintenance programme, project managers often assume a simplified solution and then will ask engineering providers to deliver to such plan. In reality, the solution is more complex. If there is a small variation (perturbation) in the planned maintenance this can cause problems later. This is known as a dominoes effect. A domino effect or chain reaction is the cumulative effect produced when one event sets off a chain of similar events. The term is best known as a mechanical effect, and is used as an analogy to a falling row of dominoes. An extension to a planned maintenance duration that deviates from the project plan can have consequences for the future entire fleet availability. To test the robustness of a project manager's plan; a Numerical Simulation called the Asset Schedule Maintenance Model (ASM) was developed. The ASM introduced small divergences in the project plan and determined the effect. This paper explores how this model can help planners in their decision making as well as predicting and allowing mitigation against the dominoes effect.

Plenary 4
Wednesday 6 September – 11.45am-12.45pm

Causality

Marloes Maathuis
ETH Zurich

Causal questions are fundamental in all parts of science. Answering such questions from non-experimental data is notoriously difficult, but there has been a lot of recent interest and progress in this field. I will discuss current approaches to this problem and outline their potential as well as their limitations. The concepts and methods will be illustrated by several examples.

RF1: Rapid-Fire Talks
Wednesday 6 September – 12.50pm-1.30pm

Vowel analysis for forensic speaker comparison

Tereza Neocleous, Zhuo Sun
University of Glasgow

In this work we model vowel formant frequencies extracted from speech recordings to assess their usefulness in the task of forensic speaker comparison. Forensic speaker comparison typically involves a disputed recording of an unknown offender (e.g. a ransom phone call) and a known sample of the suspect's voice (e.g. a police interview). The aim is to compare the speech patterns in the two samples in order to assess whether the suspect and offender recordings contain the voice of the same or different individuals. The comparison is usually made through a range of phonetic/linguistic features such as vowels and consonants, intonation, speech rate, hesitations and others. Statistical modelling of the numerical characteristics of these features is required in order to quantify the strength of evidence through a measure which takes into account both the similarity between the suspect and offender speech features and the rarity of those values within a relevant population. This measure, known in forensic statistics as the likelihood ratio, is estimated using background data available to us from a speech database. We explore models combining information from several features (i) at the modelling stage and (ii) in a post-processing step, and show that these features perform well in the evidence evaluation task.

RF1: Rapid-Fire Talks
Wednesday 6 September – 12.50pm-1.30pm

Back Propagation Gradient Based Neural Network Training Algorithms and Volatility Forecasting: Evidence from Nigeria

Shehu Usman Gulumbe, Shamsuddeen Suleiman, Yakubu Musa
Federal University Birnin Kebbi, Nigeria

Volatility forecasting has been the subject of recent empirical studies and theoretical investigation both in academia and financial markets because it is one of the primary inputs to a wide range of financial applications from risk measurement to asset and option pricing. GARCH family of models have been extensively applied in volatility forecasting, but one of their limitations is that these models produce better results in relatively stable markets and could not capture violent volatilities and fluctuations. Neural Networks (NNs) trained by Back Propagation gradient descent algorithm are known to have the capability to learn any complex approximate relationships between the inputs and the outputs with a very slow convergence rate for most practical applications. In this study, we proposed two hybrid models based on EGARCH and Recurrent Dynamic Neural Networks trained by dynamic Back Propagation gradient based training algorithms to forecast the volatility of inflation rate returns in Nigeria. The estimates of volatility obtained by an EGARCH model are fed forward to a Neural Network. The input to the first model is complemented by historical values of the other explanatory variables. The second hybrid model takes as inputs both series of the simulated data and explanatory variables. The forecasts obtained by each of those hybrid models have been compared with those of EGARCH model in terms of closeness to standard deviation which is used as a measure of the actual value of volatility. The results show that the second hybrid model trained by Bayesian Regularization algorithms gives better volatility forecasts. This model significantly improves the forecasts over the ones obtained by the EGARCH model.

RF1: Rapid-Fire Talks
Wednesday 6 September – 12.50pm-1.30pm

Two new approaches for visualising fitted models in network meta-analysis

Martin Law, Dan Jackson, Yi Yu, Navid Alam
Medical Research Council - Biostatistics Unit

Meta-analysis is a useful tool for combining evidence from multiple studies. An extension of meta-analysis, network meta-analysis, is becoming more commonly used as a way to compare multiple treatments in a single analysis.

The output of a network meta-analysis generally involves tables of treatment effect estimates, their standard errors and associated p values. Interpretation of such tables is often difficult. We present two visualisation approaches that vastly ease interpretation of the relationships between treatments in network meta-analysis. The approaches we propose are grounded in network analysis. We group treatments into “communities” by maximising the network's modularity -- a relative measure of how “good” a particular grouping is.

The first approach examines treatment effect estimates, their standard errors and associated p values. For each of these three characteristics, the treatments are grouped into communities and visualised simultaneously. The second approach examines treatment effect estimates only, and involves parametric bootstrapping based on these estimates and the covariance matrix. For each bootstrap replication, the treatments are grouped into communities, then the proportion of times each pair of treatments is in the same community is visualised using a heat map.

We illustrate our methods using a relatively large example dataset containing 22 treatments, and show how one can quickly identify which treatments have both a large effect estimate compared to standard care and are well identified. These visualisation approaches should be used by network meta-analysts to gain an increased understanding of how treatments in
a network relate to one another.

RF1: Rapid-Fire Talks
Wednesday 6 September – 12.50pm-1.30pm

Reduction of the number of animals required in pharmaceutical experiments by prediction of the compound-specific variation in oral bioavailability estimates

Nicholas Galwey, David Lugo, Leanne Cutler, Cesar Ramirez-Molina, Simon Taylor
GlaxoSmithKline

The principles of Replacement, Reduction and Refinement in the conduct of animal experiments are a constant focus in Drug Discovery. Animal experiments are only conducted when there is no suitable alternative, with pharmacokinetic characterisation a key requirement for drug optimisation. The present objective was to determine whether the number of rats used in the preliminary estimation of oral bioavailability could be reduced from 3 to 1 for each compound without affecting decisions on progression. Pharmacokinetic data on 143 compounds were available, with oral exposure summarised by the maximum blood concentration (C_{max}) and the area under the concentration-vs.-time curve (AUC). The coefficient of variation (CV) for these response variables varied widely among the compounds, from 0 to 5.23 (523%; median = 0.30) for C_{max} and from 0.02 to 11.69 (median = 0.26) for AUC. A measurement of two other drivers of oral exposure, permeability (perm) and clearance following intravenous dosing (IV), was also available for each compound, and the model '$\log_e$(response) ~ perm + IV' was therefore fitted, to see whether the magnitude of a compound's residual from this model was a predictor of its CV. For $\log_e$(AUC), in the set of compounds studied, P(CV > 0.5) = 0.225, and positive predictive value (PPV) = P(CV > 0.5|abs(residual) □ 2) = 0.659, whence relative risk (RR) = PPV/ P(CV > 0.5) = 2.92. The corresponding values for $\log_e(C_{max})$ were P(CV > 0.5) = 0.259, PPV = 0.414, RR = 1.60. On this basis it was decided that for the majority of compounds (85-90%) the first assessment of oral exposure could be made using a single rat and remain suitable for decision making. Compounds for which replicate measurements are required would be chosen on the basis of a large residual value from the model presented here, or some similar criterion.

RF1: Rapid-Fire Talks
Wednesday 6 September – 12.50pm-1.30pm

Statistical methodologies and experimental designs to aid investigating anthelmintic efficacy in cattle livestock

Johnathan Love, Louise Kelly, Hannah Lester, Ian Nanjiani, Mike Taylor, Chris Robertson
University of Strathclyde

Over the past 60 years, the use of anthelmintics has been increasingly under threat due to parasite populations (i.e. nematodes) becoming resistant to products in use. The Faecal Egg Count Reduction Test (FECRT) is the most widely used field-based method for determining drug efficacy and as an indicator of the presence of drug resistant nematodes in cattle. This test traditionally involves a parallel group design: a herd of cattle being randomised into a positive treatment and untreated control group, and their post-treatment faecal egg counts (FECs) are obtained 14 days after treatment. These measured responses can be obtained through counting techniques of various degrees of sensitivity. Afterwards the arithmetic means, T and C say, of treated and control groups are evaluated, respectively and the following percentage reduction is estimated:

$$100(1\text{-}T/C)\%$$

Published guidelines recommend using a large sample, normal approximation to evaluate a 95% confidence interval for the percentage reduction estimate above. Furthermore, alternative designs and percentage reductions, e.g. paired studies involving a positive treatment group using pre- and post-treatment counts, are also being suggested.

We have evaluated the robustness and suitability of a range of statistical methods on cattle FEC data that were obtained from large scale field studies in England. Results of the project so far have indicated that the majority of FEC data, obtained using a highly sensitive counting technique, are not normally distributed, and consequently, we recommend that confidence intervals be generated in either Bootstrapping or Bayesian frameworks. Additionally, distributions associated with the negative binomial are of better representation for these data and hence, arithmetic means should be used when calculating percentage reductions for a FECRT. We have also found for these data, via a simulation study, that Bootstrapped confidence intervals offering adequate coverage, are those associated with a positive treatment group only paired study design.

RF1: Rapid-Fire Talks
Wednesday 6 September – 12.50pm-1.30pm

Larger Control Groups in Experimentation

Marie Oldfield

To date researchers planning experiments have always lived by the mantra that ‘using equal sample sizes gives the best results’ and although unequal groups are also used, it is not the preferred method of many. Indeed Cohen and others have stated this in their academic work. However, during study planning there are other considerations that can make allocating equal sample sizes difficult such as financial cost and statistical power. My MSc investigates the hypothesis that more power can be achieved when a larger control group is used

- An extensive literature review of the area is undertaken which includes:
- Unequal Randomisation in study design
- Ethical Concerns
- Impact of sample size and variability on power and type I error rates
- Welches t test for heterogeneous variance and group sizes
- Developing the non centrality parameter when calculating sample sizes with heterogeneous variances
- The impact of cost on group sizing
- Simulated Fisher’s Test for different effect sizes and group sizes
- Simulated t-test for different effect sizes and group sizes

It was found that larger control groups may give more power to studies looking for an effect in the mid range but not to those looking for large or small effects. It was also found that the Welches test and the one and two sample t tests performed poorly with heterogeneous group sizes and variances. Recent theories in this area were explored. Recommendations are made for further research and to improve the current use of statistics and power in study design.

RF2: Rapid-Fire Talks
Wednesday 6 September – 12.50pm-1.30pm

Area estimation by double-calibration of a remote-sensed feature map to fieldwork observations via manual image interpretation

Alan Brewer
Forest Research

The paper will describe the statistical analysis of the combination of three separate spatial datasets of tree and small woodland features in the landscapes of England and Wales to produce estimates of total areas of classes of such features using calibration equations. The three spatial datasets are an automated interpretation of aerial photography covering the whole of the land area of England and Wales, a sample of 1 by 1 kilometre squares with visual interpretation of aerial photography, and a subset of these sample squares where the results of the visual interpretation were verified, augmented and corrected by ground survey.

Initial GIS processing of this data involved overlaying pairs of these spatial datasets to identify areas of intersection and non-intersection of classes of polygons in each dataset. These results were then statistically analysed using calibration models of the form $y_j=s^*\sum r_i x_{ij} +\varepsilon_j$ for areas y_j of a feature class of calibrating dataset Y in sample square j and areas x_{ij} of class i of the object dataset X in sample square j. The r_i are 'scaling-down' factors corresponding to the proportion of area of class i of dataset X that intersects with the class polygons of dataset Y across the sample, and s is a 'scaling-up' factor corresponding to the inverse of the proportion of area of the class polygons of dataset Y that intersect with those of dataset X across the sample. The first stage calibration was of the full automated interpretation dataset to the visually interpreted sample, with the second stage calibration being that of the visual interpretation to the fieldwork results.

Results will be presented for each stage of the analysis and consideration will be given to the efficiency of the method, including the question of the relative sizes of samples.

RF2: Rapid-Fire Talks
Wednesday 6 September – 12.50pm-1.30pm

What is the power to detect complex intervention effects using time-series data

Xingwu Zhou, Nicola Orsina
Karolinska Institutet

The intervention time series design is the strongest and most commonly used quasi-experimental design to assess the impact of population-based health interventions. The evaluation of the impact of public health policies on the population's health has become a major commitment for States and Communities. The validity and reliability of the longitudinal effect is likely to depend on the study design. To the best of our knowledge, the existing power studies are limited to a simple shift in level and slope in oversimplified time series modelling (McLeod and Vingilis, 2008, 2005; Zhang, Wagner, and Ross-Degnan, 2011). Therefore, our aim is to present extensive Monte Carlo simulations to evaluate the power requirements for intervention time series models under more realistic scenarios: (1) nonlinear intervention effects; (2) presence of seasonal variation; (3) varying number of time points before and after intervention; and (4) magnitude of the effect size.

The importance of conducting power calculations will be illustrated in evaluating the effects of tobacco control policies on the Swedish National Tobacco Quitline; a longstanding free service providing telephone counseling for tobacco users who want to quit the habit.

RF2: Rapid-Fire Talks
Wednesday 6 September – 12.50pm-1.30pm

Methods for Nonparametric Regression with Censored Responses

Zhou Fang, Javier Palarea
Biomathematics and Statistics Scotland

A common problem when tackling statistical modelling is how to handle records only known to be below a detection limit. Such censored data can arise in diverse situations including for instance records of chemical concentrations in water that might be constrained by the measurement technology, or DNA-based counts where the amount of DNA is measured by PCR.

Existing approaches often involve simplistic assumptions about the true values of such censored data and are prone to introduce bias and may give misleading inferences. Other approaches may only be applied to certain simple models, e.g linear regression. We aim to develop a statistically sound method to fit nonparametric regression models, including spline smoothers, to censored response variables.

We implemented two new distribution families for the popular mgcv R package for GAM analysis. The first family implements the Tobit I model relating the censored response to a latent variable, which is a function of the explanatory variables plus a normal error, through the threshold; whereas the second is similar to a restricted variant of Tobit II, introducing an additional variability parameter that separately adjusts the probability of censorship near thresholds. Both implementations allow for varying censoring thresholds across the data set and also for any mixture of left- (the case corresponding with non-detects) and right-censored data. The use of the mgcv package with these families enables access to a large variety of potential smoothers and the inclusion of simple random effects.

We present promising results in a variety of contexts, and an R package for practical implementation.

RF2: Rapid-Fire Talks
Wednesday 6 September – 12.50pm-1.30pm

Parameter Inference in Differential Equation Models using Time Warped Gradient Matching

Mu Niu, Simon Rogers, Maurizio Filippone, Dirk Husmeier
School of Mathematics and Statistics, University of Glasgow

The scientific landscape is changing, with an increasing number of traditionally qualitative disciplines becoming quantitative and adopting mathematical modelling techniques. This change is most dramatically witnessed in the life sciences. One of the most widely used modelling paradigms is based on coupled ordinary or partial differential equations (DEs). These equations are typically nonlinear, so that a closed-form solution is intractable and numerical solutions are needed. This usually does not pose any restrictions on the forward problem: given the parameters, generate data from the model. However, it does severely limit the feasibility of the statistical inference problem: given the data, estimate the underlying parameters. The reason is that for complex nonlinear systems, the log likelihood landscape is typically multimodal, calling for an iterative optimisation or sampling scheme. Since each iteration requires a numerical solution of the differential equations, the computational costs become prohibitively large.

To deal with this computational complexity, approximate methods based on gradient matching have become popular. The idea is to avoid the computationally expensive numerical solution of the DEs with an indirect approach, based on the following procedure: estimate the derivatives from the noisy data via some smoothing approach, quantify the discrepancy between these estimates and the predictions from the differential equations, and infer the model parameters based on this discrepancy. A shortcoming of this approach is the critical dependence on the smoothing scheme for function interpolation. In the talk we present a novel method that adapts an idea from manifold learning, and we demonstrate that a time warping approach aiming to homogenise intrinsic length scales can lead to a significant improvement in parameter estimation accuracy and reduced sensitivity to the smoothing scheme. We demonstrate the effectiveness of this scheme on noisy data from various dynamical systems, a biopathway, and an application to soft-tissue mechanics of the heart.

RF2: Rapid-Fire Talks
Wednesday 6 September – 12.50pm-1.30pm

Bayesian nonparametric conditional copula estimation of twin data

Fabrizio Leisen, Luciana Dalla Valle, Luca Rossini
University of Kent

Several studies on heritability in twins aim at understanding the different contribution of environmental and genetic factors to specific traits. Considering the National Merit Twin Study, our purpose is to correctly analyse the influence of the socioeconomic status on the relationship between twins' cognitive abilities. Our methodology is based on conditional copulas, which allow us to model the effect of a covariate driving the strength of dependence between the main variables. We propose a flexible Bayesian nonparametric approach for the estimation of conditional copulas, which can model any conditional copula density. Our methodology extends previous work by introducing dependence from a covariate in an infinite mixture model. Our results suggest that environmental factors are more influential in families with lower socio-economic position.

RF2: Rapid-Fire Talks
Wednesday 6 September – 12.50pm-1.30pm

Predicting extreme river discharge - comparing direct and indirect modelling strategies

Adam Butler
Biomathematics and Statistics Scotland

The quantification of extreme river discharge - flow rate - is of practical importance within the context of both high flows (which are linked to flood risk) and low flows (which are linked to water quality, and hence to freshwater ecology). Empirical data on discharge are relatively hard to collect, however, so it is common practice to routinely collect data on stage height and then to collect smaller amounts of paired data on stage and discharge. The paired data are used to infer the relationship between stage height and discharge (the "rating curve"), and the stage data are used to infer the frequency and magnitude of extreme events. Various rating curve models have been proposed; most are based upon a power-law relationship, or extensions of this.

There are also established statistical models for the analysis of extreme values - such as the Generalized Extreme Value distribution - but within the context of extreme discharge an important question arises: should the extreme value model be applied to the time series of stage height values (which are observed, but not ultimately the quantity of interest), or to the time series of discharge values (which are the key quantity of interest, but are unobserved)? In this talk we outline both possibilities, and briefly compare the performance of the two approaches within the context of simulated data. We find subtle differences between the empirical performance of the two approaches, suggesting that the choice between them will typically be non-trivial and context-dependent.

RF3: Rapid-Fire Talks
Wednesday 6 September – 12.50pm-1.30pm

Intensive blood pressure targets in clinical trial versus primary care setting

Ilyas Bakbergenuly, Elena Kulinskaya, Lisanne Gitsels, Nicholas Steel
University of East Anglia

The published SPRINT trial reported considerable survival benefits of intensive systolic blood pressure (SBP) lowering below 120 mmHg. The primary objective of our study is to investigate the survival benefits of different systolic blood pressure (SBP) targets in the US clinical setting (SPRINT) and the UK primary care setting (THIN data).

Data from the SPRINT randomised control trial and The Health Improvement Network (THIN) primary care database were used to develop survival models for longevity and adverse renal outcome (ARO, main adverse effect) at different SBP levels given treatment. To replicate the SPRINT design, we selected patients with a diagnosis of hypertension and prescription of at least one antihypertensive agent between 1995 and 2011 in THIN. Patients with SBP≤120 mmHg (analogue to intensive treatment) were matched to three controls with SBP between 120-140 mmHg (analogue to standard treatment) and SBP above 140mmHg.

The hazards of all-cause mortality and ARO associated with SBP targets were calculated by multilevel Cox’s proportional hazards regressions. Our initial analysis showed that in SPRINT, standard treatment had an increased hazard of mortality of 1.42 (1.06, 1.90) compared to intensive treatment. In contrast, in THIN, standard treatment had a reduced hazard of mortality of 0.70 (0.65, 0.76). Both in SPRINT and THIN, standard treatment was associated with a decreased hazard of ARO of 0.32 (0.22-0.46) and 0.87 (0.80-0.95), respectively

A lower SBP target was associated with increased longevity in the clinical setting, but with decreased longevity in the primary care setting. The difference in results may be due to earlier time-span of our study, and prescriptions could signify sicker patients in THIN. An intensive control of SBP may benefit a selected subgroup of patients, but it appears harmful for the broader population.

RF3: Rapid-Fire Talks
Wednesday 6 September – 12.50pm-1.30pm

Fitting Spatial Auto-Regression Models to 3-D PET Imaging Data

Tian Mou, Jian Huang, Finbarr O'Sullivan
University College Cork

Understanding of the covariance pattern in PET imaging data can contribute to improved assessment of uptake patterns that might be of clinical concern. In recent work we have been examining the use of measurements obtained in routine quality assurance as a way to develop an empirical statistical description of PET scans used in clinical practice. Part of the analysis involves consideration of spatial auto-regressions. We model each voxel value as linear combination of its neighbours' ones. The first and second order neighbours are considered. Unlike least-squares auto-regressive analysis used for AR models in standard time series settings, for our models, we find it necessary to estimate the spatial coefficients by minimising the difference between the estimated autocorrelation pattern of the data and the pattern implied by the model. A nonlinear least-squares algorithm, implemented in R, has been developed for the minimisation. To overcome the potential problem of the large data size, we propose to use a random sample of voxels and their corresponding neighbours as part of the analysis. The fitted model is applied to all the voxels and the residuals are analysed to guide the selection of structure of the spatial model, including its extent. Simulation studies and asymptotic analysis are performed to validate the proposed methodology. The results show that proposed algorithm can estimate the model parameters consistently. Secondly, the performance of the proposed method is demonstrated on normalized PET data. The method can be adapted to provide very practical mechanisms for routinely simulating PET images with noise characteristics associated with quantitative PET studies.

RF3: Rapid-Fire Talks
Wednesday 6 September – 12.50pm-1.30pm

Some simple designs for censored survival trials

Alan Kimber, Maria Konstantinou, Stefanie Biedermann
University of Southampton

In a two-armed randomised trial where the aim is to estimate the treatment effect, it is clearly best to have an equal number of subjects in each arm trial, isn't it? Not necessarily if the response is the time to an event and may be censored. In this talk we see that locally optimal designs are easy to find in the case of possibly censored exponential event times, that these designs have interesting robustness properties, both in terms of misspecification of parameter values and of misspecification of event-time distribution. These designs also turn out to be nearly optimal in some situations if we prefer to use a semi-parametric Cox proportional hazards model rather than a fully parametric one.

RF3: Rapid-Fire Talks
Wednesday 6 September – 12.50pm-1.30pm

Conveying risks in transplantation to patients

Kate Martin, James Neuberger, Dave Collett, Rachel Johnson, Ron Stratton
NHS Blood and Transplant

There are currently over 6000 patients on the waiting list for an organ transplant in the UK, with just over 4600 solid organ transplants taking place last year. This disparity raises an important question for many patients - when will I get transplanted? Patients are also interested in the risk involved in receiving a transplant, the chance of receiving an organ from a high risk donor, and the risk of a transmitted infection. Providing patients with the information to answer these questions using a public website would enable them to make an informed decision about proceeding with a transplant, by weighing up the risks involved themselves.

A clear and accessible way to present this is to show *what will happen to 100 people like me if I go on the transplant waiting list*? To answer this we would need to account for relevant patient characteristics as well as the organ to be transplanted. Our initial focus is on patients who need a liver transplant.

Robust data from the UK Transplant Registry on 2,838 patients registered for a liver only transplant and 2,017 liver transplants between 1 April 2007 and 31 March 2011 were used to analyse patient outcomes. Post-registration and post-transplantation outcomes at one, three and five years were broken down by patient's primary liver disease and a binary measure of disease severity. Outcomes are presented in graphic form with 100 figurines representing the proportion of patients who are alive or have died. The risks involved and the cause of failure or death are presented in a similar graphic.

A public website for use by patients and their families gives them an avenue to information previously not available for them, informing them of the balance between the potential benefit and also the potential risks involved in receiving a liver transplant.

RF3: Rapid-Fire Talks
Wednesday 6 September – 12.50pm-1.30pm

Can testing clinical significance reduce false positive rates in clinical trials?

Theophile Bigirumurame, Adetayo Kasim
Durham University

A lot has been said over the years about the limitation of null hypothesis significance testing, particularly with respect to false positive rate among the published findings (Ioannidis 2005). What is also clear is that most of the problem is how null hypothesis significance testing is used rather than the statistical approach itself (Robinson & Wainer 2001). Another concern is the mismatch between power calculation and the analysis of trial data. Typically, sample size calculation for randomised control trial is based on a minimum clinically significance difference that should support both statistical significance and clinical findings from the trial. The use of minimum clinical significance difference in the hypothesis formulation for superiority trial is similar in principle to the concept of non-inferiority or equivalence trial. However, most clinical trials are analysed testing zero clinical difference. Since minimum clinical significance difference has been pre-defined for power calculation, it is natural to incorporate this value in both the testing and interpreting the importance of findings from clinical trial.

We reviewed a set of 50 publications (25 with binary outcome, and 25 with survival time outcome). 16% of 25 published trials with binary outcome that were statistically significant, were also clinically significant based on the minimal clinical significance risk difference used for their power calculation. 24% of the 25 published trials with survival time outcome that were statistically significant were also clinically significant based on the minimum clinical hazard ratio used for their power calculation.

These results seem to suggest that accounting for the minimum clinical significance in the analysis and interpretation of trial data may reduce false positive rate. However, a systematic review is needed to critically appraise the impact of the current practice on false positive rate in published trials with significant findings.

RF3: Rapid-Fire Talks
Wednesday 6 September – 12.50pm-1.30pm

Statistical Analysis on Longitudinal Data for the prospect of introducing a Joint Pain Advisor into the Primary Care setting of the NHS

Hira Naveed, Rosie McNiece, Mike Hurley, Andrew Walker
Kingston University and Health Innovation Network

Chronic conditions such as Osteoarthritis in the older population are the leading causes of disability and are a continuing financial strain on the NHS. The objective of this paper was to evaluate the efficacy of a newly developed initiative, the Joint Pain Advisor. Which is a service that aims to give advice to individuals on how to manage their hip and knee joint pain. The participants of the study were seen at three time points and their responses were measured repeatedly over this time. The Hip Osteoarthritis Outcome Score and the Knee Osteoarthritis Outcome Score (HOOS and KOOS) were used to measure pain and physical function. The study design was aimed at the older population (n=498). A series of statistical tests and models, such as the ANCOVA and the Repeated Measures Model were used on the longitudinal data, this involved independent variable analysis as well as including an interaction term. This helped with assessing the effect of the service on pain and physical function. The analysis was also adjusted to avoid bias, especially due to missing values. Results concluded a significant interaction between age and gender, BMI showed a significant effect on outcome scores, the cohort had a reduced amount of pain and less difficulties with physical functioning and mental wellbeing also improved drastically. The findings in this paper provide substantial evidence of the positive effect the service had on participants, both mentally and physically. In addition, feedback received from the cohort after completing the programme was also very positive.

RF4: Rapid-Fire Talks
Wednesday 6 September – 12.50pm-1.30pm

Comparing Methods for Early Detection Systems for Seasonal and Pandemic Influenza

Mugrin Almugrin, Chris Robertson, Alison Gray
University of Strathclyde

The aim of this work is to investigate established methods for early detection of seasonal and pandemic influenza, then to develop a new method for routine use. We investigate daily data for influenza-like illness (ILI) General Practice (GP) consultations, collected from all 14 spatially located Health Boards (HBs) in Scotland. We use these real data to generalise their applicability to any number of geographic areas. The National Health Service (NHS) provided data on ILI consultation data from 2009 to 2015.

Our work is extending the Weekly Cases Ratio (WCR) method to develop an automatic system to raise an alarm when there is a sudden increase in ILI cases. The WCR method uses two terms: 1) the value of WCR, defined in week w as WCR(w)=ILI GP consultation rate in week w divided by the corresponding rate in week w-1, and 2) the number of health boards N(HB) which report an increase in ILI cases in week w compared to week w-1 (i.e. WCR(w)>1).

We initially used the above Scottish data, then extended this to more than 14 HBs, by simulating data with similar structure but from a different number of HBs. We used a constant rate of consultations to determine a joint null distribution for (WCR, N(HB)) to use in a hypothesis testing approach to detect a rise in ILI cases. Using more than 3 million simulations of each data set, we found a relationship in all cases between WCR and N(HB), then modelled the relationship between each of the mean (μ) and standard deviation (σ) of WCR with N(HB) in each dataset to find a general equation for use with any number of HBs. Further work will compare our approach with the Moving Epidemic and Cumulative Sum methods.

Some results of this work will be presented.

RF4: Rapid-Fire Talks
Wednesday 6 September – 12.50pm-1.30pm

An integrated machine learning approach of feature selection and classification of long intergenic non-coding RNA in heart failure patients

Mintu Nath, Simon Romaine, Jamie Timmons, Christopher P Nelson, John R Thompson, Adrian A Voors, Nilesh J Samani
University of Leicester and on behalf of the BIOSTAT Consortium

The diverse role of non-coding RNAs in human pathophysiology is well-recognised yet remains incompletely understood. Identifying long intergenic non-coding RNAs (lincRNAs), associated with the survivability of heart failure patients, could elucidate novel regulatory mechanisms and functional relevance of lincRNAs, and predicting the risk in such patients would benefit clinically. We adopted an integrated machine learning approach to analysing the expression profile of 5369 lincRNAs, generated with the Affymetrix Human Transcriptome Array 2.0, from 944 age and sex-matched heart failure patients. Among these patients, 318 died while 626 survived during the two-year follow-up period. We analysed the log2-transformed normalised expression data in three stages. Firstly, we conducted the CUR matrix decomposition on the full data, prepared a dataset retaining 1666 most influential lincRNAs with leverage scores greater than the mean. Secondly, using regularisation, nearest shrunken centroids, sparsity and a combination of these strategies, we conducted dimension reduction, feature selection and classification of 75% of data (training data) with a set of ten different models capturing both linear and non-linear class boundaries. Finally, we identified three dissimilar (Jaccard dissimilarity index: 0.80-0.88) and better performing models – elastic net penalised regression, high dimensional regularised discriminant analysis and flexible discriminant analysis – and employed the gradient boosting method to develop a prediction model as an ensemble of these three models. Results showed estimates in the training (75%) and testing (25%) datasets of the ensemble prediction model for the parameters (training, testing) area under the curve (0.81, 0.65), sensitivity (0.74, 0.71), specificity (0.76, 0.56) and accuracy (0.75, 0.63) were better or comparable with individual models, fitted on datasets with or without the CUR matrix decomposition. We conclude that the proposed integrated machine learning approach, along with the enhanced computational efficiency, identified lincRNAs of biological interests and improved the prediction based on patients' expression profiles.

RF4: Rapid-Fire Talks
Wednesday 6 September – 12.50pm-1.30pm

A Joint modelling approach to access the association between child and adult HiV infections in Kenya, having adjusted for covariates.

Elvis Karanja
University of Nairobi

Recent studies have adopted a joint modelling approach as a more robust technique in studying outcomes of interest simultaneously especially when the interest is in the association between two dependent variables. This has been necessitated by the fact that modelling such outcomes separately often leads to biased inferences due to existing possible correlations especially in medical studies. This paper focuses on establishing if there exists a correlation between child and adult HiV infections measured for each county in Kenya, while adjusting for several predictors such as coverage of anti-retroviral therapy (ART) in each county, the number of adults and children in need of ART amongst other variables. We obtain HiV data for each county from the Kenya government open data website for the year 2014 and visualize on each county the HiV infections on the Kenyan map. High infection incidences are observed for counties located in Nyanza province of Kenya. We further jointly model the two outcomes of interest using the linear mixed models approach for repeated measures to capture the correlation between the two outcomes for each county. Results indicate the infections are indeed correlated with significant predictors such as ART coverage, Adults and Children in need of ART as well as number of people undergoing testing voluntarily.

RF4: Rapid-Fire Talks
Wednesday 6 September – 12.50pm-1.30pm

Identifying harmful medications in pregnancy: use of a double False Discovery Rate method to adjust for multiplicity

Alana Cavadino, Joan Morris, David Prieto-Merino
Queen Mary University of London

Continued surveillance of medication use in pregnancy is essential to detect new teratogens. Medications in the same Anatomical Chemical Therapeutic (ATC) classes may work in similar ways, and we aimed to use this information to improve the detection of teratogens using congenital anomaly surveillance data.

EUROmediCAT is a network of European congenital anomaly registries. Data on 15,058 malformed fetuses with first trimester medication exposures from 1995-2011 were extracted from the EUROmediCAT database. For each anomaly and medication combination the proportion of cases of an anomaly in women taking the medication was compared to the proportion of that anomaly in all other women in the dataset, resulting in 28,765 combinations (55 anomalies x 523 medications). To adjust for multiplicity a "single" adjustment to control the false discovery rate (FDR) was initially used across all medication-anomaly combinations. We refined this methodology by applying a "double" FDR procedure incorporating an additional step to consider groupings of medications according to their ATC codes. The Australian classification system for prescribing medicines in pregnancy was used to independently identify “high risk” medications. The number of “high risk” medications and the total number of signals identified by single and double FDR were compared.

A higher proportion of the resulting set of signals were “high risk” for double FDR compared to single FDR. Double FDR also identified more “high risk” medications overall, for a range of FDR cut-offs and comparable effective workloads. Evaluation of signal detection methods in congenital anomaly data is difficult due to the lack of a "old standard" to classify risks for medications in pregnancy according to anomaly (the Australian classification system does not identify the specific anomalies associated with each high risk medication). We recommended, however, that double FDR be used in future routine signal detection analyses of congenital anomaly data.

RF4: Rapid-Fire Talks
Wednesday 6 September – 12.50pm-1.30pm

A comparison of seven random-effects models for meta-analysis that estimate the summary odds ratio

Dan Jackson, Martin Law, Theo Stijnen, Wolfgang Viechtbauer, Ian White
MRC

Comparative trials that provide binary outcome data are commonly pooled in systematic reviews and meta-analyses. This type of data can be presented as a series of two by two tables. The pooled odds ratio is often presented as the outcome of primary interest in the resulting meta-analysis. This talk will examine the use of seven models for random-effects meta-analysis that have been proposed for this purpose. The first of these models is the conventional one that uses normal within-study approximations. The other six models are generalised linear mixed models that have the potential to provide more accurate inference. Empirical, simulation based and analytical investigations will be presented. We conclude that generalised linear mixed models can result in better statistical inference than the conventional random-effects model, but also that these models present their own issues and difficulties. In particular, one of the seven models that we investigate is found to perform poorly. This finding has serious implications for models for meta-analysis that include fixed study effects, such as those typically used to perform individual patient data meta-analyses.

RF4: Rapid-Fire Talks
Wednesday 6 September – 12.50pm-1.30pm

Missing Data Analysis Under Censoring in Data Obtained from Healthcare Organisations

Kate Pyper, Chris Robertson, Michael Eddleston, Alastair Rushworth
University of Strathclyde

Missing data are commonly found in real life statistical practice. There are three well documented types of missingness: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). Where the missing data are MCAR and MAR, the missing data process is termed ignorable, and can be factored out of the likelihood. Where the data are MNAR, the missing data mechanism is non-ignorable and the missing data mechanism must be known in order to make reasonable inference.

Data requests often result in datasets where cells are suppressed due to small values. Therefore it is only known that a censored value is within a certain range of values, or specifically less than a particular value. This naturally means that the missing data depend on the value of the variable of interest, and is therefore non-ignorable. The work was carried out in an attempt to make good parameter estimation in relation to hospital admissions due to poisoning per hospital per month. This dataset was obtained from the Information Services Division of the NHS, whose policy is to suppress any data cells with values less than five.

In this situation, common missing data methods do not generally apply. This work attempts to use what is known about the missing values in order to build a robust parameter estimation process. Bayesian data augmentation is used to estimate the parameters, where at each update the missing data are imputed with draws from the truncated distribution, where the parameters are estimated by a linear model.

This method has been compared to other common methods for parameter estimation under missing data as part of a simulation study. The results of this have indicated that making use of the additional information produces estimates with better properties than those produced using other missing data methods.

RF5: Rapid-Fire Talks
Wednesday 6 September – 12.50pm-1.30pm

Evidence Based Analysis for Investment Decision Support

Petros Gousias, Lesley Walls, Matthew Revie, Nicolas Jego, Athena Zitrou
University of Strathclyde & Scottish Water

Water Quality: Evidence Based Analysis for Investment Decision Support

Drinking water regulations require high water quality to be delivered at the customer tap. The journey from the water source to the customer tap is a system exposed to hazards that can impact water quality. Bacteriological growth is a critical hazard that may first enter the water system at the source of the catchment. This risk is managed through the process controls within water treatment works. But since treated water can be stored in service reservoirs, designed to add capacity to the water distribution system, there are new opportunities for bacteriological growth before delivery to the customer.

An analysis of the likelihood of bacteriological non-compliance across the water system is being conducted to help inform the investment planning decisions by Scottish Water.

The analysis is based upon empirical data, including flow cytometry records, for the Scottish water system collected from January 2013 to December 2016. The explanatory factors include water quality data, catchment area data, weather data, asset characteristics and treatment process data. Interesting insights include that chlorine levels, year of construction of sites and design capacity, amongst others, are found to be key risk factors for particular measures of bacteriological growth. Moreover, bacteriological risk is associated with the cleaning frequency of tanks, providing evidence of the efficacy of investment in this operational action.

A system-wide model is being developed within a mixed effect Generalised Linear Model framework to support additional analysis. For example, to provide a national view of the relative risks in the annual investment planning period based upon the recent observational data, and to explore the predicted effects of investment choices on the risk reduction in water quality non-compliance.

RF5: Rapid-Fire Talks
Wednesday 6 September – 12.50pm-1.30pm

Understanding the Different Approaches of Measuring Owner Occupiers' Housing Costs

Arthur Eidukas
Office for National Statistics

Owner occupiers' housing costs (OOH) are the costs of owning and living in one's own home. This is distinct from the price of a house, which is not consumed in the same way as other goods and services. OOH, therefore, is not necessarily an easily relatable concept; however, it is a major part of consumer expenditure, and measuring these costs is an important and challenging part of consumer price statistics. We currently publish the Consumer Prices Index including OOH (CPIH). CPIH uses our preferred method – rental equivalence – for measuring these costs.

There are multiple ways to measure OOH, each with a different methodological and conceptual underpinning. In this analysis, we produce several different experimental measures. Our goal is to provide clarity on how these measures differ from each other and to show the main drivers behind changes in OOH for each method.

We consider comparisons of growth trends and growth rates for each method. We also explore the sub-components of each index to determine the main drivers of growth. This provides a better understanding of how each component affects the overall measure of OOH. Finally, we look at how the OOH indices would affect a comprehensive measure of price growth. We compile special aggregate versions of CPI-H, which we use to see how the underlying components of OOH affect inflation more generally, as measured by these aggregates.

Our results suggest that, unlike our rental equivalence measure, the experimental measures are often driven by asset prices or movements in the interest rate which should not be a part of consumer price inflation.

RF5: Rapid-Fire Talks
Wednesday 6 September – 12.50pm-1.30pm

What can we learn from ONS Foreign Direct Investment statistics?

Shane O'Connor, Sami Hamroush, Yanitsa Petkova
Office for National Statistics

This presentation will outline the latest statistics on Foreign Direct Investment (FDI) in terms of positions, earnings and implied rates of return.

The analysis covers:

- FDI by continent: The analysis also uses normalised data to focus on FDI with all EU member states in 2015 in addition to identifying changes among the UK's main European FDI partners since 2007.
- FDI by industry: Counterfactual analyses are used to assess the impact that changing implied rates of return may have had on FDI statistics. There is also a focus on mining & quarrying, financial & insurance services and manufacturing.
- Distributional analysis: Assessing the impact of earnings by size of direct investment.
- Exchange rate effects: Estimating the possible impact that changes to the sterling exchange rate may have had on FDI statistics between 2011 and 2015, as well as over the first three-quarters of 2016.

Findings:

The fall in UK FDI credits is partly explained by changes in the implied rates of return on UK FDI assets, which have been falling since 2011; in contrast, returns on liabilities have been relatively more resilient.

The values of FDI assets in the majority of EU member states were below medium-term averages in 2015.

Falling FDI credits from mining & quarrying has been an important factor for outward earnings, while increases in FDI debits from UK's financial & insurance industries have provided notable positive contributions to debits.

The UK's largest 25 companies in terms of total overseas investment were found to be the main driver of the downward trend recorded in FDI credits since 2011.

Exchange rate movements were not the main determinant of the value of UK assets between 2011 and 2015; however, the depreciation of sterling over the first three-quarters of 2016 appears to have had a positive impact.

RF5: Rapid-Fire Talks
Wednesday 6 September – 12.50pm-1.30pm

GSS Harmonisation Programme: Past, Present and Future

Ian Sidney, Becki Aquiina
Office for National Statistics

The Harmonisation Team works across the Government Statistical Services (GSS) to produce harmonised definitions, questions and outputs. These harmonised principles are reused across national statistics allowing users to more easily compare data from different sources and makes our statistics easier to understand.

The GSS Harmonisation Programme has developed a range of harmonised definitions, questions and outputs for a range of key social topics over the past ten years or so. However it is now looking at developing business harmonised principles as well.

The Bean Review, EUROSTATs Framework Regulation Integrating Business Statistics (FRIBS) and the requirements of the UK Government's Digital by Default initiative are providing an ideal opportunity to harmonising business survey questions and definitions. The Office for National Statistics (ONS) is aiming to move all business surveys from paper to an electronic data collection mode (EDC). The EDC programme gives rise to a unique opportunity to review all ONS business survey questions and to develop harmonised business survey definitions, questions and outputs and through a process of rationalisation reduce the number of surveys wherever possible. This has the potential to reduce cost of collection and respondent burden. It will potentially improve the quality, timeliness and accuracy of data and could increase response rates from certain groups.

This presentation will focus on what has been achieved to date with social and business harmonisation, how we work across the GSS to develop and encourage the use of the harmonised principles, how key programmes such as the 2021 Census are instrumental to the harmonised programme and how you can help to develop and promote these harmonised principles.

RF5: Rapid-Fire Talks
Wednesday 6 September – 12.50pm-1.30pm

Estimating the number of rooms and bedrooms in the 2021 Census: An alternative approach using Valuation Office Agency data

Natalie Shorten
Office for National Statistics

In May 2016, the ONS published its updated view on the topics proposed for inclusion in the 2021 Census. The 2021 Census Topic Consultation identified there was a medium user need for data about number of rooms and number of bedrooms. These were collected in 2011 using two separate questions on the census questionnaire.

The ONS decided it was not appropriate to continue to ask two questions designed to meet a single information need, therefore a commitment was made to explore use of administrative data from the Valuation Office Agency (VOA) to gather data on number of rooms and meet Eurostat requirements.

ONS has undertaken analysis on the quality of Valuation Office Agency data in comparison to the 2011 Census data. In June 2017, we will publish rates for number of rooms, number of bedrooms and occupancy rating based on the commitments made following the topic consultation.

We will highlight the opportunities and challenges of using VOA as an alternative data source for producing the number of rooms and number of bedrooms, with consideration to the following:

- The potential for producing more frequent outputs from the VOA dataset
- Definitional differences between VOA and Census for the number of rooms / bedrooms
- Difficulties in aligning VOA data with traditional census definitions of a ‘household’

We will conclude with a discussion about the feedback received from users and our approach to making the decision on the ‘number of rooms’ question for inclusion in the 2021 Census.

RF5: Rapid-Fire Talks
Wednesday 6 September – 12.50pm-1.30pm

Trumpets and orchestras: some thoughts on optimising public value

Elizabeth Fraser
Scottish Government

In what has been described by some as a post-truth era, the position and value of statistics, both official and otherwise, becomes rather more interesting (in a good way). We have access to a plethora of information, although some ‘information’ is more useful than others, and too much information can sometimes be more hindrance than help. Truth may be contingent or contested, requiring us to make sense of conflicting evidence. This tends to be particularly the case in the most interesting areas for public policy, where real life is complex but dwelling on complexity may stifle progress.

Highlighting what it is our statistics are telling us about the world we live in is important, be it confirmation of what we already know or suspect, or illuminating something new. So how do we realise the full potential of the statistics we produce?

Drawing on examples across a range of policy areas, this paper explores how we can look at our statistics in different ways in order to draw out what is most valuable for a range of audiences. We often focus on the big hit at point of publication, but may perhaps neglect, or underestimate, the longer term cumulative impact of the body of knowledge we generate, resulting in a biased appreciation of the public value of the resource we produce. The paper also aims to encourage statisticians and users of statistics to become more curious about the impact statistics have, or could have, on society, and how they can contribute to enhancing the public value of those statistics.

RF6: Rapid-Fire Talks
Wednesday 6 September – 12.50pm-1.30pm

Statistical simulation of longitudinal HIV viral load trajectories

Maia Lesosky, Tracy Glass
University of Cape Town

Longitudinal HIV viral load is a critical measure of short and long term treatment efficacy in HIV-infected individuals. As a direct and responsive measure of replicating HIV in the body it represents a best case biomarker. However, HIV viral load is also complex from a statistical perspective. It has high variability, where log order magnitude changes can happen in a matter of days, it is subject to censoring, due to different assay limits of detection and it is subject to non-trivial magnitude dependent, measurement error. These characteristics make viral load trajectories difficult to model, and difficult to simulate.

In order to evaluate different viral load monitoring strategies for routine care in resource limited settings, we required a large set of frequently measured HIV viral load, data that is typically not available outside of randomised controlled trials. Using contributed data from a clinical trial in South Africa and routinely collected laboratory data from the national laboratory service to parameterise the model, we developed a novel individual Monte-Carlo simulation model of longitudinal HIV viral load that accurately replicates observed viral load trajectories in individuals initiating antiretroviral therapy, over 18 months of follow up.

The simulation model, sensitivity analysis, and some discussion of specific modelling choices particularly around values below the limit of detection are discussed. The parametric and semi-parametric assumptions modelled for the suppression and rebound trajectories are discussed in the context of making assumptions about longitudinal data based on repeated cross sectional observations.

RF6: Rapid-Fire Talks
Wednesday 6 September – 12.50pm-1.30pm

Logistic modeling of mothers' attitude towards immunization in Nigeria

Akeyede Imam, Saleh Musa Ibrahim
Federal University Lafia

Immunization programme has had a major impact on the health status of the world population, by preventing many cases of infectious disease. Efficient vaccine storage and handling is a key component of immunization programme. It is a shared responsibility from the time the vaccine is manufactured until it is administered. Thousands of children were dying and some are disable as a result of some common diseases which are measles, polio, tetanus, whooping cough, tuberculosis etc, hence, there is need to sensitize the parents towards the immunization of their children. This study therefore, aimed at fitting binary logistic model that will describe the pattern of mother's attitude towards immunisation in Nigeria in order to identify the factors that responsible for the attitudes in Nigeria.. Questionnaires were administered to 5000 women in some states in Nigeria to elicit relevant information regarding their general attitudes to child's vaccination. Binary logistic regression was used to analyse the data obtained on Demographic and other factors considered. Results from analyses showed that mothers locality, place of vaccination, mothers educational status, age at vaccination, spouses educational status, mothers' religious believes as well as mothers' age group, child's age at birth are all positively associated with attitudes of mothers towards vaccination. Further results finally revealed that donation of gift items to mothers serves as positive inducement towards improving the attitudes of mothers towards immunization of their children.

Keywords: Logistic Modeling, Immunisation, Demographic-Characteristics

RF6: Rapid-Fire Talks
Wednesday 6 September – 12.50pm-1.30pm

Opportunities and challenges of analysing UK Biobank data

Francesca Chappell, David Dickie, Caroline McHutchison, Joanna Wardlaw
University of Edinburgh

Objectives. To highlight issues around analysis of the UK Biobank dataset.

Background. The UK Biobank dataset (http://www.ukbiobank.ac.uk/about-biobank-uk/) is a resource established by the Wellcome Trust, available to researchers based anywhere. Over 500,000 UK participants contributed extensive health-related data, giving a unique opportunity to investigate predictors of disease.

Data were collected from people aged 40-69, initial assessments were from 2006–2010 and follow-up is ongoing. UK Biobank also comprises the largest health imaging study – so far ~5000 of 100,000 individuals have undergone MRI scanning.

UK Biobank is an enormously important resource for investigating causes of disease. However, researchers need to act with due diligence in both the analysis and interpretation of results.

Challenges include:

- The UK Biobank Demographic dataset is 7GB of data – complex analyses require extra planning and this does not include imaging or genetic data
- UK Biobank participants are healthier than the general UK population with a lower 5 year mortality rate, so inference to UK general population is difficult
- Missing data can be extensive – for example, fluid intelligence has 67% missing data at baseline, and data are probably not MAR
- Counterintuitive results – fractional anisotropy measures brain white matter structural integrity and decreases with age, white matter lesions increase with age. However, fractional anisotropy in white matter lesions increases with age – suggesting that the lesions of older versus younger people are healthier. Younger people also have slightly higher deprivation scores – studies of age effects may be confounded.

Conclusions. Inference from the UK Biobank sample to the general UK population requires care – mechanisms of missingness and impact of the self-selected sampling will affect results. Unexpected health differences between age groups exist.

References

Ganna A, Ingelsson E. 5 year mortality predictors in 498 103 UK Biobank participants: a prospective population-based study. Lancet 2015;386:533-540.

RF6: Rapid-Fire Talks
Wednesday 6 September – 12.50pm-1.30pm

Estimating trends in population-level HIV viral suppression from routine laboratory data

Elton Mukonda, Maia Lesosky, Landon Myer, Nei-Yuan Hsiao
Division of Epidemiology/University of Cape Town

Global targets for antiretroviral therapy (ART) in HIV treatment programmes call for 90% of those on ART to achieve sustained viral suppression by the year 2020. Low and middle income countries have only recently developed capacity for digital collection and storage of routine data. Thus, the lack of availability of quality population based programmatic data is a major hindrance to the measurement of progress of intervention programs in these countries. HIV viral load monitoring is the method of choice to evaluate treatment effectiveness and transmission risk in HIV-infected individuals, and has undergone recent scale up in South Africa.

We analysed 1,1 million routinely collected viral load results from the National Health Laboratory Service (NHLS), Western Cape, South Africa during the period 2008-2015. The background scale up in services went from an average of 5840 tests per month to 20,707. During this period routine monitoring guidelines changed from biannual to annual, and treatment eligibility was broadened. We estimate trends in the sample and population viral suppression (defined as VL < 1000 copies/mL). In addition to describing and estimating the time trends in the sample by facility and age related strata, we estimate the population viral suppression rate by adjusting for the South African Census estimated numbers of HIV-infected individuals in the province.

Preliminary results suggest there is considerable variation between facilities with the overall sample viral suppression proportion in the largest 10 facilities ranging from 81%-89%, while smaller facilities range from 33%-91%. Overall viral suppression has been increasing, from 82% to 84%. The estimated HIV-infected population viral suppression in the Western Cape has increased from 15% to 37% over the 7 years of data.

Though significant progress has been made in the light of large increases in numbers of patients on therapy, there remain challenges in reaching the final 90-90-90 target.

RF6: Rapid-Fire Talks
Wednesday 6 September – 12.50pm-1.30pm

Using follow-up panels as a sampling frame – what are the considerations of efficient sampling?

Andrea Lacey, Cathy Jones
Office for National Statistics

Since 2011, the Labour Force Survey (LFS) has included a question to ask respondents for consent to follow-up surveys after their last interview. This follow-up panel provides a sampling frame of around 20,000 individuals per quarter, making it one of the largest probability sample interview panels in the UK. With pressure on ONS to reduce data collection costs and reduce respondent burden in our surveys, this sampling frame allows the use of information collected on the LFS to design an efficient sample of the target survey population and use cheaper data collection modes such as telephone interviewing. However, given that the follow-up panel has suffered from attrition between the first and final interview, and consent to follow-up may not be random, could this efficiency be at a price of introducing bias?

ONS has commenced fieldwork for two surveys: the 2016/17 Adult Education Survey (AES), a survey of 15,000 individuals, and the Opinions Survey (OPN), a survey of 2000 individuals, using the LFS follow-up panel as a sampling frame. This presentation will take the AES and OPN as examples to outline the opportunities and challenges of using this sampling frame.

RF6: Rapid-Fire Talks
Wednesday 6 September – 12.50pm-1.30pm

Automatic Balancing of UK Supply and Use Tables in Current Prices: 2011 to 2014

Neil Parkin, Joshua Abramsky, Craig McLaren, Christopher Quickfall
Office for National Statistics

The unbalanced UK supply and use tables (SUT) in current prices are several years of measurements of output, taxes and subsidies, imports and exports, consumption, capital formation, compensation, and profits. Balancing improves the quality of estimates of GDP by adjusting the unbalanced SUT so that they satisfy economic identities: total supply of each product equals total use of that product, and gross value added (GVA) production for each industry equals GVA income for that industry. The balanced SUT are vital inputs to dynamic models of the UK macro-economy built and used by, for example, the Bank of England and the Treasury. Balancing is currently carried out by human experts, and is a difficult and time consuming process. A model has been built of the manual balancing of SUT. The modeled data are for 112 different products and industries and the years 2011 to 2014, the unbalanced tables contain 460,000 measurements. The model has been programmed as a constrained weighted least-squares problem, and is solved by a standard ONS laptop in less than two minutes. The model can be extended to improve its quality and scope. Extra knowledge put into the model by human experts will improve the quality of balanced SUT, and will effectively document the human decisions that were made to balance SUT. The scope of the model can be widened in many ways, for example: to impute unmeasured parts of the unbalanced SUT; to measure any biases in data, should they be present; to account for auto-correlations; to cover more years of SUT; to balance sectors of the accounts. The automatic method will be used also to balance valuation matrices, a key part of producing SUT in previous years prices and double-deflated estimates of GDP.

RF6: Rapid-Fire Talks
Wednesday 6 September – 12.50pm-1.30pm

What have been the effects of tobacco control policies on advice-seeking at Swedish National Tobacco Quitline?

Nicola Orsini, Xingw Zhou, Alessio Crippa, Rosaria Galanti
Department of Public Health Sciences, Karolinska Institutet

Knowledge about the series of phone calls received by a smoking cessation quitline in response to different interventions aiming at reducing tobacco smoking is currently lacking. Aim of this study is to examine the possible effect of four types of policies on the calling rates to the Swedish smoking cessation quitline: a campaign on passive smoking (Jan 2001); placing larger text warnings on cigarette packs (Sept 2002); banning smoking from restaurants (Jun 2005); and a 10% tax increase (Jan 2012). We used 16-years of monthly data collected between January 1999 to December 2014 (192 months) counting a total of 162,978 phone calls. Upon definition of four pre-post intervention intervals, we used intervention time series ARIMA (Auto-Regressive Integrated Moving Average) models where the outcome was defined as calling rates expressed per 100,000 smokers. Rate ratio (RR) at 6 months after intervention together with a 95% confidence interval (CI) were derived from the model. The campaign on passive smoking on Jan 2001 was associated with a 85% higher calling rate (95% CI=1.13-3.04). Larger text warnings on cigarette packs in Sept 2002 conferred a 53% increment in the calling rate (95% CI=1.20-1.94). Smoking-free restaurants was associated with a significant 11% (95% CI=1.00-1.1.23) higher calling rate. The 10% tobacco tax increase in Jan 2012 had no significant effect on the calling rate (RR=0.98, 95% CI=0.82-1.15). Within an overall decreasing trend in the population of smokers in Sweden, we were able to detect differential effects of smoking policies on the calling rates to the quitline, the most effective being the campaign on passive smoking and the larger text warnings signs on the cigarette packs.

6.1 Contributed - Medical statistics: Methods in epidemiology and public health
Wednesday 6 September – 2.30pm-3.30pm

Estimating treatment effects with optimal inverse probability weighting

Michele Santacatterina, Matteo Bottai
Unit of Biostatistics - Karolinska Institutet

The average treatment effect is defined as the difference in the expectation of an outcome of interest between treatment groups. It is a popular summary measure, but when the reatment assignment is not random, it may not be directly interpretable. This is because potentially relevant factors may be unequally distributed across the treatment groups, and any observed differences in the outcome may not be accounted for by the treatment alone. Inverse probability weighting is often used to adjust for this potential unbalance. This method, however, is known to yield erratic and inefficient inference when outlying weights are present. Different approaches have been proposed to alleviate this limitation. These frequently entail introducing simplifying assumptions in the model for estimating the probability of being treated. While these approaches generally reduce the variability in the weights, and the consequent sampling variability of the weighted estimators, they can also introduce substantial bias. We present optimal inverse probability weighting, which minimizes the bias of the weighted estimator of the average treatment effect for any specified level of its standard error. The optimal weights are defined as the solution to a nonlinear constrained optimization problem. The method is evaluated in a simulation study and applied in the assessment of the timing of treatment initiation in individuals infected by the human immunodeficiency virus. The simulation study suggests that optimal inverse probability weighting has some desirable properties, such as:
(1) it provides an estimated ATE with minimum bias while controlling for precision;
(2) it allows the researcher to search for a suitable balance between bias and variance directly;
(3) it can be implemented with available R packages, such as ``ipoptr'' and ``nloptr''; and
(4) it maintains all its properties regardless of the chosen estimator for the ATE.

6.1 Contributed - Medical statistics: Methods in epidemiology and public health
Wednesday 6 September – 2.30pm-3.30pm

Hierarchical group testing with multiplex assays in heterogeneous populations

Chris Bilder, Joshua Tebbs, Christopher McMahan
University of Nebraska-Lincoln

Testing individuals for infectious diseases is important for disease surveillance and for ensuring the safety of blood donations. When faced with questions on how to test as many individuals as possible and still operate within budget limits, public health officials often use group testing (pooled testing) with multiplex assays (multiple-disease tests). The testing process works by amalgamating specimens from individuals (e.g., blood, urine, or saliva) into groups and then applying a multiplex assay to each group. For low disease prevalence settings, the majority of these groups will test negatively for all diseases; thus, greatly reducing the number of tests needed in comparison to individual testing with single-disease assays. For those groups that test positively for at least one disease, algorithms have been developed to retest sub-groups and/or individuals in order to distinguish the positive individuals from those who are negative. The purpose of this presentation is to provide a first-of-its-kind algorithm that incorporates individual risk information into the retesting process for multiplex assays. Through simulation and application, we show that this new algorithm reduces the number of tests needed in comparison to those procedures that do not include individual risk information, while also maintaining sensitivity and specificity levels.

6.1 Contributed - Medical statistics: Methods in epidemiology and public health
Wednesday 6 September – 2.30pm-3.30pm

Value of Information: Sensitivity Analysis and Research Design in Bayesian Evidence Synthesis

Christopher Jackson, Anne Presanis, Stefano Conti, Daniela De Angelis
MRC Biostatistics Unit, University of Cambridge

Suppose we have a Bayesian model which combines evidence from several different sources. We want to know which model parameters most affect the estimate or decision from the model, or which of the parameter uncertainties drive the decision uncertainty. Furthermore we want to prioritise what further data should be collected. These questions can be addressed by Value of Information (VoI) analysis, in which we estimate expected reductions in loss from learning specific parameters or collecting data of a given design. We describe the theory and practice of VoI for Bayesian evidence synthesis, using and extending ideas from health economics, computer modelling and Bayesian design. The methods are general to a range of decision problems including point estimation and choices between discrete actions.

We apply the methods to a model for estimating prevalence of HIV infection, combining indirect information from several surveys, registers and expert beliefs. This analysis shows which parameters contribute most of the uncertainty about each prevalence estimate, and provides the expected improvements in precision from collecting specific amounts of additional data. We also discuss the application of these methods to a model for the health impacts of transport policies and scenarios.

6.2 Contributed - Official statistics and public policy: Maternity and births
Wednesday 6 September – 2.30pm-3.30pm

Variations in numbers of births by day of the week in relation to onset of labour and mode of giving birth, England 2005-2014

Peter Martin, Mario Cortina Borja, Nirupa Dattani, Gill Harper, Mary Newburn, Miranda Dodwell, Alison Macfarlane
City, University of London

Background: Maternity care has to be available 24 hours a day, seven days a week. It is known that obstetric intervention can influence the timing of birth but there are few studies investigating how the day of birth is associated with the onset of labour and the mode of giving birth at a national level.

Methods: As part of a project to analyse the timing of birth and its outcomes we linked data from birth registration, birth notification, and Maternity Hospital Episode Statistics, to analyse data about over 5 million births in NHS maternity units in England from 2005 to 2014. To analyse the association between the day of the week of birth and the onset of labour and mode of giving birth by gestational age, we fitted negative binomial regression models to the daily frequency of births, controlling for seasonal cycles of births as well as trends over time.

Results: The frequencies of birth varied considerably by onset of labour and mode of giving birth. Births after spontaneous onset and spontaneous delivery are slightly more likely to occur on weekdays than at weekends and on public holidays, and around 7 % less likely to occur on Christmas than on an average day. Elective caesarean births are concentrated onto weekdays. Births after Induced labours are more likely to occur on Tuesdays to Saturdays and on days before a public holiday period, than on Sundays, Mondays and during or just after a public holiday.

Conclusion: Frequencies of birth vary by care pathway, and these patterns have implications for midwifery and medical staffing. The timing of interventions is partly determined by the working rhythms of maternity units. We explore possible reasons why term births without interventions occur less frequently at weekends and on holidays than during the working week.

6.2 Contributed - Official statistics and public policy: Maternity and births
Wednesday 6 September – 2.30pm-3.30pm

Insights on Third Party Linkage and Data Quality from Quality Assuring Linked Birth Registrations and HES Delivery Records in England 2005 to 2014

Gill Harper
City University

Linked records from birth registration from the Office for National Statistics (ONS) for all births that in England 2005-2014 were linked to Maternity Hospital Episode Statistics (HES) delivery records by NHS Digital, using an in-house algorithm and maternal identifiers. Quality assurance prepared this linked dataset for analysis for the 'birth-timing and outcomes' project which is studying the daily, weekly and yearly cycles of birth-timings and their implications for child and maternal health. The linked dataset contained duplicate and incorrect linkages. Quality assurance aimed to select a single correctly linked HES record with the maximum of delivery information for each ONS birth record.

The method categorised linkage types to identify which links should be preserved and which should be discarded. Comparison of common baby data items and the availability of valid values in other key fields were used to inform this. Singleton and multiple births were processed separately due to the increased complexity of multiple birth linkage.

After quality assurance, 95% of all singleton births and 93% of all multiple births were each correctly linked to a single HES delivery record. Most of the discarded linked records were duplicate HES records.

Differences between correct and incorrect links were found, often relating to data quality issues. Some missed links were also discovered, raising questions about the quality of the linkage.

The quality assurance process for the linked dataset has provided more knowledge about the quality of the data and of the linkage, than could be obtained from cleaning each dataset in isolation. The availability of identifiers in all three datasets supported exploration and confirmation of true links and errors. Other research projects using linked large routinely collected administrative datasets from trusted third parties should not assume that the linked datasets are error-free or optimised for their analysis.

6.2 Contributed - Official statistics and public policy: Maternity and births
Wednesday 6 September – 2.30pm-3.30pm

‘A time to be born’?

Alison Macfarlane, Peter Martin, Mario Cortina-Borja, Nirupa Jobanputra (formerly Dattani), Gill Harper, Mary Newburn, Rod Gibson
City, University of London

Background

It has long been established that numbers of spontaneous births varied by time of day and were higher at night than during the day and data showing this were cited by statisticians, notably Adolphe Quetelet and Austin Bradford Hill. The extent to which the rise in obstetric intervention has modified this pattern has not so far been documented at a national level in England and Wales, as national data about the time of day of birth were not available before 2005.

Methods

In a previous project, data recorded when a baby’s birth is registered by parents were linked to the data recorded when a birth is notified to the NHS, including the time of birth and this linkage was mainstreamed by ONS. These previously linked data about births from 2005 to 2014 were now linked to data about care at birth recorded in the Hospital Episode Statistics for England and corresponding data for Wales. Births were analysed by hour of the day and day of the week by gestational age, onset of labour and mode of birth and results for singleton births in England will be described.

Results

About half of the births occurred spontaneously after spontaneous onset of labour. Their numbers were highest between 1am and 7am and peaked around 4am. For other births following spontaneous onset of labour, instrumental births were more likely to occur between 9am and 5pm and emergency caesareans were more likely between 9am and midnight. Births following induced labours peaked around midnight, regardless of mode of birth. In contrast, the majority of planned caesareans occurred between 9am and noon.

Interpretation

These patterns have implications for the staffing of maternity units. Although some births can be scheduled to take place in day time hours, many babies will still be born at night.

6.3 Contributed - Environmental/Spatial statistics: Environmental Epidemiology
Wednesday 6 September – 2.30pm-3.30pm

Bayesian hierarchical modelling of social genetic effects in livestock disease transmission

Osvaldo Anacleto, Andrea Wilson, Santiago Cabaleiro
Roslin Institute, University of Edinburgh

Despite significant advances in statistical methods for infectious disease data, current stochastic epidemic models ignore genetic heterogeneity in host infectivity, which is the propensity of an infected individual to transmit pathogens to susceptible individuals. Variation in this social interaction trait can lead to the common superspreading phenomenon in disease outbreaks, where a minority of highly infected hosts are responsible for transmitting the majority of infections. Geneticists have long been interested in exploiting heritable variation in infectivity to reduce disease severity in livestock production. However, to date it is not known to what extent infectivity and superspreading are genetically controlled.

We present a novel stochastic transmission model which, by combining individual-level Poisson processes with bivariate random effects, can fully capture genetic variation in infectivity. Using simulation data, we show that not only can this Bayesian model accurately estimate heritable variation in both infectivity and the propensity to be infected, but it also can identify parents more likely to generate offspring that are disease superspreaders. An application based on a large-scale fish infection experiment shows, for the very first time, that genetics does indeed contribute to variation in infectivity and therefore affects the spread of diseases.

6.3 Contributed - Environmental/Spatial statistics: Environmental Epidemiology
Wednesday 6 September – 2.30pm-3.30pm

Preliminary investigation of the influences on antimicrobial resistance

Katie Stewart, Louise Matthews, Marian Scott, Dirk Husmeier, Colin McCowan
University of Glasgow

Antibiotic resistance is an important threat to human health, as well as to livestock health and welfare. This has become evident in the last few years with the increased occurrence of resistant infections within humans such as MRSA, *C. difficile* and *E. coli*. This paper reports on progress investigating the availability of antibiotic usage and resistance data and what can be determined in relation to the resistance problem from these data and their links to environmental conditions.

Urinary Tract Infections (UTIs) are common in the community and widely treated with antibiotics. Patients with UTI are at higher risk of developing *E. coli* bacteraemia, which can cause severe complications and death. There are three main antibiotics which are used to treat UTIs, namely Nitrofurantoin, Trimethoprim, and Cefalexin, with a further three including Amoxicillin, Ciprofloxacin and Co-amoxiclav. Our work will focus on this selection of six antibiotics, initially using data for Scotland and England GP prescriptions from 2015-2016 and 2014-2016 respectively.

The broad aims of the study include determining whether antibiotic use in humans is associated with antimicrobial resistance, and what patterns of antibiotic use or resistance exist in the wider community. Statistical tools used include data visualisation and spatio-temporal smoothing methods in order to produce a series of maps showing the underlying distributions of prescriptions and other factors.

6.3 Contributed - Environmental/Spatial statistics: Environmental Epidemiology
Wednesday 6 September – 2.30pm-3.30pm

Divide and conquer: Partitioning mosquito biting heterogeneity and identifying malaria hotspots for intervention

Su Yun Kang, Donal Bisanzio, David Smith
University of Oxford

Malaria is a leading cause of childhood mortality, responsible for 438,000 deaths each year. A large portion of heterogeneity in the intensity of malaria transmission over time and space can be attributed to seasonality, individual household attractiveness, environmental noise, and measurement error. With the ability to identify households which act as hotspots for sustaining a major portion of malaria transmission we can increase the efficiency of control interventions.

This study focuses on mosquito count data from entomological surveillance conducted between October 2011 and September 2016 at three study sites in Uganda. A total of 330 households were involved in the surveillance. The intensity of malaria transmission varies considerably across the three sites. Using a Bayesian zero-inflated negative binomial model and a prior distribution for seasonal signals, we partitioned the heterogeneity in mosquito abundance into household biting propensities, seasonality, and environmental and measurement noise. We also conducted hotspot analysis using the Getis-Ord statistic for all three sites to identify malaria hotspots among the households. For each study site, we contrasted household biting propensities in different scenarios – dry vs. rainy seasons; before the enrolment of indoor residual spraying vs. after; during the first half period of the surveillance vs. the second half period of the surveillance.

Our work on partitioning the heterogeneity in mosquito abundance provides an understanding of heterogeneity in malaria exposure and offers a critical appraisal of the possibility of targeting interventions at households with the most mosquitoes, which differ depending on various environments and various levels of endemicity. Focusing malaria control efforts on households who contribute disproportionally to malaria transmission could achieve community protection by eliminating transmission in a relatively small fraction of human hosts.

6.4 Contributed - Social statistics: Socio-demographics and Inequalities
Wednesday 6 September – 2.30pm-3.30pm

Multidimensionality of longitudinal data: Unlocking the age-happiness puzzle

Ning Li
Australian Mathematical Sciences Institute

In longitudinal analysis of social economic data, sometimes the explanatory variables that are statistically significant in OLS regressions in cross-section or pooled data become insignificant after controlling for individual fixed effects. This phenomenon was observed in the study of the relationship between age and happiness. Majority studies in the literature of happiness research found that happiness is U-shaped in age. The U shape, however, was challenged by a qualitatively different finding in other studies that control for individual fixed personalities. These studies found that happiness decreases with age. The discrepancy in estimated age-happiness relationship between the U shape and the decline pattern was known as the age-happiness mystery. In this paper, I shall points out that OLS regressions based on cross-section data, which resulted in the U shape, reflect the average difference in happiness across birth cohorts. In contrast, regressions controlling for individual fixed effects reflect the change in happiness over time within individuals. For the first time in the literature, the co-existence of a cross-section U shape and a longitudinally decline pattern in the relationship between age and happiness is established. Using data from the Household, Income and Labour Dynamics in Australia (HILDA) Survey, this paper explains the exact meaning of fixed-effects regression of happiness on age, gives insight into the age-happiness puzzle, and raises the awareness of multidimensionality of longitudinal data.

6.4 Contributed - Social statistics: Socio-demographics and Inequalities
Wednesday 6 September – 2.30pm-3.30pm

Gender Pay Gap Elements in the UK and Scotland

Wendy Olsen, Wasel Bin Shadat, Sook Kim, Giuseppe Maio, Min Zhang
University of Manchester Department of Social Statistics

Pay gaps are notoriously slippery measures, because there are options for the choice of denominator. The pay gap fell during the last decade in the UK. Decomposition is used to show that pressures upward on the pay outcome are compensated for, in part, by pressures downward. The interpretive wording can be endowments+unexplained; or characteristics vs. coefficients; or drivers vs. protective factors. We give examples to illustrate the 'drivers vs. protective factors' interpretation of a gender pay gap in the UK. In the Scottish case a smaller sample N leads us to question whether its drivers differ from those in the UK as a whole.

We consider four issues – the units in which to measure the gap components; the nuances of wording; how the clustering of household members in the underlying Understanding Society survey data may affect our estimates of regional pay gaps (non-exchangeability); and how we measure change over time in the drivers. Change over time in the drivers and protective factors can be discerned even when there is no net change over time. Bootstrapping is used to refine and validate each of the findings.

Specifically, nonparametric bootstrapping under exchangeability shows London leading the UK pay gap regional rankings. Parametric and wild bootstrapping offer wise choices which allow the causal factors to be integrated into the model. London has drivers which differ from the rest of the UK. We modify the model to allow for non-exchangeability of individuals. A multilevel model with households as level 2 and regions as Level 3 can better illustrate the pay gap by region for the UK.

We acknowledge the inputs of David Bayliss, Vanessa Gash, Hein Heuvelman, Pierre Walthery, Leen Vandecasteele, Anna Ritchie, staff at Close the Gap Glasgow, staff at the Government Equalities Office and Arek Wisniowski.

6.4 Contributed - Social statistics: Socio-demographics and Inequalities
Wednesday 6 September – 2.30pm-3.30pm

Measuring Change Over Time in Socio-Economic Deprivation and Health in an urban context. The case study of the city of Genoa

Stefano Landi, Enrico Ivaldi, Angela Testi
University of Genoa

The existence of an inverse association between socioeconomic status and health status is well established in the literature. Difference in health due to socioeconomic are unfair because could be avoided with better health and social policies. The relation between socioeconomic factors and health inequality may be proved at the individual level, or at the geographic-area level. In this paper we follow the second stream of literature, i.e studies on deprivation relating the state of disadvantage suffered by an individual, with the living conditions of the area where the individual resides.

The aim of this study is to measure relative disadvantage in an urban contest and compare the results with health oucomes over time.

A census based deprivation index have been devoloped using three different aggregation methods (Additive, Mazziotta-Pareto Index and Pena Distance method). A cluster analyses has been run to group the areas in homogeneous cluster with respect to the Deprivation score.

In particular have been analysed the association between a deprivation index and standardised (premature) mortality ratios. Secondly have been showed how the standard mortality ratio change according to deprivation clusters.

Deprivation inequalities in Genoa are still present but globally they are lowering over years. In regard to health status Premature Mortality has improved over years. Moreover the number of areas and population living in a deprived area is lowering. Despite this positives, our case study confirms the literature results that socioeconomic conditions affect health status deeply.

6.5 Contributed - Methods and theory: Big Data
Wednesday 6 September – 2.30pm-3.30pm

Modelling High Dimensional Volatilities by Common Factors

Jiazhu Pan
University of Strathclyde

We consider a framework for modelling conditional variance (volatility) of a high dimensional time series by common factors. We propose to estimate the factor loading space by a space generated by the orthonormal eigenvectors corresponding to the r largest eigenvalues of a matrix based on the observations, and develop the asymptotic theory on the proposed estimation method based on the empirical process theory. Some novel asymptotic results on empirical processes constructed from nonstationary random sequences, which pave the way for the main results, are presented.

We also discuss the consistency of our eigenanalysis estimation for the loading matrix as both the cross sectional dimension and the sample size go to infinity. We further illustrate the methodology using both simulated and real data examples.

6.5 Contributed - Methods and theory: Big Data
Wednesday 6 September – 2.30pm-3.30pm

Approximate posterior inference for Markov random fields with discrete states

Matt Moores, Anthony Pettitt, Kerrie Mengersen
University of Warwick

There are many approaches to Bayesian computation with intractable likelihoods, including the exchange algorithm and approximate Bayesian computation (ABC). A serious drawback of these algorithms is that they do not scale well for models with a large state space. Markov random fields, such as the Ising/Potts model and exponential random graph model (ERGM), are particularly challenging because the number of discrete variables increases linearly with the size of the image or graph. The likelihood of these models cannot be computed directly, due to the presence of an intractable normalising constant. In this context, it is necessary to employ algorithms that provide a suitable compromise between accuracy and computational cost.

Bayesian indirect likelihood (BIL) is a class of methods that approximate the likelihood function using a surrogate model. This model can be trained using a pre-computation step, utilising massively parallel hardware to simulate auxiliary variables. We review various types of surrogate model that can be used in BIL. In the case of the Potts model, we introduce a parametric approximation to the score function that incorporates its known properties, such as heteroskedasticity and critical temperature. We demonstrate this method on 2D satellite remote sensing and 3D computed tomography (CT) images. We achieve a hundredfold improvement in the elapsed runtime, compared to the exchange algorithm or ABC. Our algorithm has been implemented in the R package “bayesImageS,” which is available from CRAN.

6.5 Contributed - Methods and theory: Big Data
Wednesday 6 September – 2.30pm-3.30pm

A Bayesian model selection approach for identifying differentially expressed transcripts from RNA sequencing data

Panagiotis Papastamoulis, Magnus Rattray
University of Manchester

Recent advances in molecular biology allow the quantification of the transcriptome and scoring transcripts as differentially or equally expressed between two biological conditions. Although these two tasks are closely linked, the available inference methods treat them separately: a primary model is used to estimate expression and its output is post processed by using a differential expression model. This talk will present the recent model of Papastamoulis and Rattray [1], [2], where both issues are simultaneously addressed by proposing the joint estimation of expression levels and differential expression: the unknown relative abundance of each transcript can either be equal or not between two conditions. A hierarchical Bayesian model builds on the BitSeq framework and the posterior distribution of transcript expression and differential expression is inferred by using Markov chain Monte Carlo sampling. It is shown that the model proposed enjoys conjugacy for fixed dimension variables; thus the full conditional distributions are analytically derived. Two samplers are constructed, a reversible jump Markov chain Monte Carlo sampler and a collapsed Gibbs sampler, and the latter is found to perform better. A cluster representation of the aligned reads to the transcriptome is introduced, allowing parallel estimation of the marginal posterior distribution of subsets of transcripts under reasonable computing time. The algorithm proposed is benchmarked against alternative methods by replicating independent large scale simulation studies and applied to real RNA sequencing data.

References

[1] P Papastamoulis and M Rattray (2017). A Bayesian model selection approach for identifying differentially expressed
transcripts from RNA sequencing data. Journal of the Royal Statistical Society: Series C (Applied Statistics).
doi: 10.1111/rssc.12213
[2] P Papastamoulis and M Rattray (2017). Bayesian estimation of Differential transcript usage from RNA-Seq data. arXiv:1701.03095

6.6 Contributed - Medical statistics: Statistics in Biology
Wednesday 6 September – 2.30pm-3.30pm

When the microbiome meets the metabolome: An integrative longitudinal analysis of characteristic metabolic profiles concurrent with gut microbiota changes

Takoua Jendoubi, Panagiotis A Vorkas, Timothy M D Ebbels, Robert C Glen
Imperial College London

Metabonomics time-course experiments provide the opportunity to observe the evolution of metabolic profiles in response to internal and external stimuli. Along with other omic longitudinal profiling technologies, these techniques have great potential to complement the analysis of complex relations between variations across diverse omic variables and provide unique insights into the underlying biology of the system. However, many statistical methods currently used to analyse short time-series omic data are i) prone to overfitting or ii) do not take into account the experimental design or iii) do not make full use of the multivariate information intrinsic to the data. The model we propose is an attempt to i) overcome overfitting by using a weakly informative Bayesian model, ii) capture experimental design conditions by means of a mixed-effects model and iii) model the temporal dependencies using an auto-regressive component and interdependencies between variables by augmenting the mixed-effects model with a CAR prior.

We present our methodology in the context of a randomized controlled trial in a rat model. In this study, comprehensive metabolic phenotyping of the effect of metformin was performed, by monitoring the longitudinal metabolic variations in the plasma of healthy Wistar rats. The analysis was complemented with 16S rRNA gene sequencing in order to observe concurrent changes in the gut microbiome and attempt to identify potential blood biomarkers of putative changes in the gut microbiota. Results show that mild alterations observed within the gut microbiome are associated with a cascade of changes involving the host organism metabolome and gut microbiome, possibly supporting the hypothesis of an inter-level feedback loop.

6.6 Contributed - Medical statistics: Statistics in Biology
Wednesday 6 September – 2.30pm-3.30pm

Parameter Inference in the Pulmonary Blood Circulation

Mihaela Paun, Mansoor Haider, Nicholas Hill, Mette Olufsen, Muhammad Qureshi, Theodore Papamarkou, Dirk Husmeier
University of Glasgow

In my work I have focused on inferring parameters in two models:

- A partial differential equations model of pulmonary circulation in humans (two-sided: arterial and venous circulation)
- A partial differential equations model coupled with a Windkessel model in mice (arterial circulation)

The model takes some parameter values and aims to mimic the behaviour of the pulmonary haemodynamics under normal physiological and pathological conditions. This is of medical relevance as it allows monitoring the lung disease progression. Flow measurements come from MRI scans at the inlet of the main pulmonary artery, and the pressure is currently measured in an invasive way for the human patients. Hence, the rationale behind this project is to create a model able to predict pressure, therefore avoiding any invasive procedures be performed on patients.

The model solves a system of nonlinear partial differential equations and outputs predicted blood flow and pressure at different locations along the vessels. The model input consists of various bio-parameters characterising fluid dynamics and blood vessels geometry; these parameters are not measured in-vivo, so they need to be inferred indirectly from the measured flow and pressure.

Exploratory analysis has revealed that the objective function, the residual sum of squares between the measured signal and the generated signal from the model is a convex function, at least in a 2-dimensional space. Therefore, nonlinear optimization has been employed, and for simulated data, where the parameters are known, the method learns the parameters well. Currently, I am extending the work to real mice data, where it appears that there are parameter unidentifiability issues. In order to deal with this, I have taken the analysis further and Monte Carlo Markov Chain methods have been applied. In my talk I will present results based on this.

6.6 Contributed - Medical statistics: Statistics in Biology
Wednesday 6 September – 2.30pm-3.30pm

Statistical inference of the drivers of collective cell movement

Dirk Husmeier, Elaine Ferguson, Jason Matthiopoulos, Robert Insall
University of Glasgow

Collective movements of eukaryotic cells are essential for the occurrence of many major biological processes, including tissue development, wound healing, the immune response, cancer cell invasion and metastasis. The majority of the mechanisms proposed as drivers of cell movement invoke the process of chemotaxis, whereby cells bias their movement in response to gradients in the concentration of certain chemicals (chemoattractants) in their environment. Alternatively, gradients can form through local depletion of a widely-produced chemical. Several recent studies have revealed cases where cells move in response to gradients that they have created themselves by depletion of a chemoattractant, sparking new interest in the area of self-generated gradients.

To understand and influence collective cell movement, we need to be able to identify and quantify the contribution of their different underlying mechanisms. In the present work, we define a set of nine candidate models, formulated as advection–diffusion–reaction partial differential equations, to describe various alternative mechanisms that may drive cell movement. This includes random-walk type diffusion, self-generated gradients in a chemo-attractant that the cells deplete locally, chemical interactions between the cells, and various saturation processes due to limited binding capacity of the cell membranes. We fit these models to high-resolution microscopy time series from two different cell types: Dictyostelium discoideum and human melanoma. The model parameters are sampled from the posterior distribution using the delayed rejection adaptive Metropolis algorithm (DRAM). The “best” partial differential equation model is selected using the Widely Applicable Information Criterion (WAIC). By identifying the most likely drivers and mechanisms of collective cell movement, the statistical inference methods that we present shed more light on the relevance of self-induced gradients, provide a guide for future experimental work and may suggest new medical intervention strategies.

This work was carried out at Softmech, the EPSRC Centre for Multiscale Soft Tissue Mechanics.

6.7 Invited - Prize Winners: Research Students Conference 2017
Wednesday 6 September – 2.30pm-3.30pm

Spatio-temporal Modelling for Road Accident Hotspot Prediction

Joe Matthews
Newcastle University

Road traffic safety is a global issue, with around 1.3 million people killed on the world's roads each year, leading the United Nations to declare the decade 2011-2020 to be a "Decade of Action on Road Safety" with the goal of halving global traffic deaths by the end of this decade. The task of improving road safety mainly falls at a local level to road safety practitioners working for local authorities/councils, who attempt to improve safety by employing road safety countermeasures/treatments (speed cameras/traffic lights etc). The process of selecting when and where best to implement these schemes is heavily reliant on accident data. However, reliable accident data is usually scarce, leading to these decisions being made on very small numbers of observations. This opens the door for statistical phenomena such as "Regression To the Mean" (RTM) to mislead decisions, leading to inefficient allocation of resources, and potentially deadly stretches of road left untreated.

Here we propose a novel Bayesian hierarchical structure to model accident rates and predict future collision counts at road accident sites across a network, allowing for a proactive allocation of resources and removing the need for large numbers of accident to occur before action is taken. The model proposed makes use of the characteristic features of a site as covariates, to form a regression model known as a safety performance function. This allows us to identify extraneous effects (such as RTM) present in our data, even when few observations are available. We use geographically weighted regression to observe the spatially varying effects of these covariates on accident rates, as well as modelling macroscopic changes in accident rates across the network using a kernel density smoother. Finally we explore the possibility of a seasonal effect on accident rates on a monthly scale, and model this using a conditional autoregressive structure.

6.7 Invited - Prize Winners: Research Students Conference 2017
Wednesday 6 September – 2.30pm-3.30pm

Sequential Monte Carlo Methods for Epidemic Data

Jessica Welding, Peter Neal
Lancaster University

Epidemics often occur rapidly, with new data being obtained daily. Due to the frequently severe social and economic consequences of an outbreak, this is a field of research that benefits greatly from on-line inference. This provides the motivation for developing a Sequential Monte Carlo algorithm that can provide real-time analysis of outbreak data. The underlying idea being that we can generate initial samples at time T and repeatedly update them as new data is obtained.

In this talk we will discuss the construction of an SMC algorithm that can be applied to outbreak data. This algorithm will use a combination of reweighting and resampling to update the current samples as new information is received. The algorithm constructed is shown to be comparable in estimation capabilities to the current 'gold-standard' using Markov chain Monte Carlo methods, with the additional advantage of being easily parallelized, thereby allowing for a reduction in computation time.

6.7 Invited - Prize Winners: Research Students Conference 2017
Wednesday 6 September – 2.30pm-3.30pm

Development of new pain in older Irish adults: a latent class analysis of health risk factors

Aoife O'Neill, Kieran O'Sullivan, Mary O'Keeffe, Ailish Hannigan, Cathal Walsh, Helen Purtill
University of Limerick

Objectives: Pain significantly restricts the quality of life and well-being of older adults. With our increasingly aging population it is important to examine whether differing classes of health risk factors can predict the development of pain in older adults.

Methods: Latent class analysis (LCA) provides a probability model-based approach to identifying underlying subgroups in a population based on some measured characteristics. In this study, LCA was used to identify different health risk classes in people aged over 50, from The Irish Longitudinal Study on Aging (TILDA), who reported not being troubled by pain at Wave 1 and participated in the follow-up interviews two years later at Wave 2 (N=4458). Associations between the health risk classes and new pain were examined using logistic regression, adjusting for socio-demographic variables.

Results: At follow-up (Wave 2), 797 (17.9%) of participants reported being troubled by pain. Four latent classes were identified based on eleven potential health and lifestyle risk factors at Wave 1. These classes were characterised as 'Low Risk', 'Physical Health Risk', 'Mental Health Risk' and 'High Risk'. The Low Risk class accounted for over half the sample (51.18%), while the High Risk class represented 7.8% of the sample. The High Risk class was more likely to develop new pain compared to the Low Risk class (OR= 3.16, 95% CI= 2.40, 4.16). These results add to existing data in other populations supporting the role of a range of health and lifestyle risk factors which increase the risk of developing pain. These findings have important implications for the identification, and potential moderation, of these risk factors.

6.8 Contributed - Industry and commerce: Assessing Innovation, Investment and Individualised Inference
Wednesday 6 September – 2.30pm-3.30pm

A new method to calculate a corporate innovation index

Gloria Gheno, Massimo Garbuio
Free University of Bozen-Bolzano

Background

Measuring the level of innovation of a company is essential in a global and highly competitive economy, innovation is essential for the development of a business. The production of the best product in the market, indeed, does not guarantee the long-term survival of the company in an evolving economic climate like the current one. In such scenario, companies with good sensing processes are able to determine and to discover the best opportunities for innovation in the industry.

Objectives

The objective of this study is to find a scientific method to obtain an index ordering the 50 most innovative Australian companies based on their commitment to the development of sensing capabilities, using publicly available data and therefore without costs. We use our index to examine and to identify specific factors necessary for innovation. It can be applied to companies of other Countries so companies from different geographical regions can be compared.

Methods

To determine the behavior of a specific company, we analyse if the companies invest in corporate ventures, accelerators, incubators, innovation labs and internal co-working spaces. To determine how much each type of investment weights on our index, we use our new statistical method which is based on factor analysis and principal component. Calculated the index for each company, we build one for each sector. At first our index is applied to 50 Australian companies and subsequently to 30 companies of Singapore.

Conclusion

In an evolving economy like the present one, it is important for a company to be able to remain competitive on the market and thus always to innovate. Measuring innovation of a company, therefore, becomes essential to remain competitive and therefore we propose an index which calculates a value utilized for corporate policies. Being built with a strict statistical-mathematical method, it is replicable.

6.8 Contributed - Industry and commerce: Assessing Innovation, Investment and Individualised Inference
Wednesday 6 September – 2.30pm-3.30pm

Smart Beta and Empirical Alpha Representation

G Charles-Cadogan
University of Leicester, Division of Finance, School of Business

The problem posed is one in which a portfolio manager wants to increase the returns on her portfolio, relative to a benchmark or market portfolio, in order to improve her “capture ratio” (a measure of an investment’s compound return relative to a benchmark’s compound return). To do so [s]he alters the betas of the portfolio in anticipation of market movements, and augments that portfolio with hedge factors. This includes but is not limited to revising asset allocation or readjusting portfolio weights within an asset class. In other words, altered betas represent the manager’s portfolio strategy. Conceptually, the allocation of assets in the local benchmark is “fixed” but hedge factors are stochastic–at least for so called “portable alpha”. These “smart beta” strategies contemplate “active decisions to identify the specific factor(s) to target, and to define the factor(s), the selection universe, the weighting method, and the re-balancing rule. These decisions are made at the outset of the investment process, rather than throughout the process,” Jacobs and Levy (Journal of Portfolio Management, 2014). This paper’s contribution to portfolio theory and behavioural finance, and the gargantuan market timing literature, stems from its reconciliation of active portfolio management with efficient markets when portfolio strategy or investment style is unobservable. It employs asymptotic theory to identify an empirical portfolio alpha process with dynamic portfolio adjustments that reflect managerial strategy via martingale system equations that portend algorithmic trading. Additionally, it proves that the measurable sets for portfolio manager market timing ability are much larger than those proffered in the extant literature which tests for timing ability via statistical significance of convex payoff structure(s). Accordingly, we propose a new and simple test for market timing ability based on the spectral circle induced by a behavioural transformation of the hedge factor matrix.

6.8 Contributed - Industry and commerce: Assessing Innovation, Investment and Individualised Inference

Wednesday 6 September – 2.30pm-3.30pm

Individualised Inference in Practice

Tim Drye
Data Analysts User Group

Our objectives have been to follow the proposals for Keli Liu, Xiao-Li Meng individualised inference and apply them in practice. These proposals suggest that available data is first micro-sampled to identify a collection of sampling points that are similar to a particular entity of interest. This collection of sampling points become the reference set within which to generate a model for the individual entity.

We have applied these approaches in three diverse applications, to analysis and benchmark the activities of individual GP practices within NHS England, to provide a basis for identify relevant previous horse races, that help with predictions of the outcome of a particular race, to select a group of consumers who help to build a customer service model, as an individual presents themselves within a call centre environment.

In each context we show how you might traditionally generate a global model, with appropriate levels, and then apply this to the individual environment, and contrast this with the outcomes that arise when using this individualised approach.

7.1 Invited - Health economic evaluation studies - how to deal with missing data Wednesday 6 September – 3.40pm-5pm

Missing data in health economic evaluation

Andrew Briggs
Glasgow University

This presentation will set the scene for the other presentations in the session. Drawing on examples from past experience, Andrew will explore the reasons why missing data may be particularly problematic for health economic evaluation. These reasons are more practical than fundamental in that the issue of missing data does not require different methods so much as a better understanding of why the problems of missing data in economic evaluation may be more acute than in other settings. The aim will be to try and give both a historical perspective to the problem, while acknowledging recent advances and also pointing to the need for future methods development.

7.1 Invited - Health economic evaluation studies - how to deal with missing data
Wednesday 6 September – 3.40pm-5pm

Practical sensitivity analysis framework for health economic analyses with missing data

James Carpenter, Alexina Mason, Manuel Gomes, Richard Grieve
University College London & London School of Hygiene and Tropical Medicine

Health economic studies with missing data are increasingly using approaches such as multiple imputation that assume that the data are Missing at Random (MAR). This untestable assumption is often questionable. This is because - even given the observed data - the probability that data are missing may reflect the true, unobserved outcomes, such as patients' true health status.

In these cases, methodological guidelines call for the development of practical, accessible, approaches for exploring the robustness of conclusions to departures from MAR.

In this talk, we propose a Bayesian framework for CEA where outcome or cost data are missing. Our framework includes a practical, accessible approach to sensitivity analysis that allows the analyst to draw on expert opinion.

We illustrate the framework in a CEA comparing an endovascular strategy (eEVAR) with open repair for patients with ruptured abdominal aortic aneurysm, and demonstrate software for prior elicitation, allowing ready implementation of this approach.

Reference:

Mason AJ, Gomes M, Grieve R, Ulug P, Powell, JT and Carpenter JR (2017) Development of a practical approach to expert elicitation for randomised controlled trials with missing health outcomes: application to the IMPROVE trial. *Clinical Trials*, **14**, 357-367.

7.1 Invited - Health economic evaluation studies - how to deal with missing data
Wednesday 6 September – 3.40pm-5pm

Full Bayesian models to handle missing values in cost-effectiveness analysis from individual level data

Gianluca Baio
University College London

Cost-Effectiveness Analyses require the modelling of a relatively complex structure of relationships between some measure of clinical effects (e.g. QALYs) and the associated costs. Standard normality assumptions about the distribution underlying such variables are sometimes unsatisfactory, especially when they are characterised by a spike at a boundary. For this reason, the application of established methods (eg multiple imputation) may become complicated and are in practice not very popular in health economics.

We propose a full Bayesian approach to handling missing data for cost-effectiveness analysis, with a general framework able to deal with correlation across the two outcomes, as well as the presence of mixtures of components (e.g. patients with utility identically equal to 1, or costs identically equal to 0). The framework is presented using a case study to describe its structure and economic implications.

7.2 Invited - Devolution in Official Statistics
Wednesday 6 September – 3.40pm-5pm

Siobhan Carey
Chief Statistician, Northern Ireland

Siobhan will describe the opportunities that arise from being small and fleet of foot. Northern Ireland has been innovating in data collection with the development of electronic data collection for business through the Integrated Business Survey and is well advanced in developing the infrastructure for online household surveys. As the only part of the UK to have a land border with the EU post Brexit this session will also describe how NISRA responded to the need for data on cross border movement of people and goods since Article 55 was triggered.

John Morris
Welsh Government

John will talk about the challenge of cross-UK comparability as policies diverge looking at experiences in health and education. He will also talk about the challenge of meeting user need to deliver detailed and good quality statistics to support a devolved policy agenda.

Roger Halliday
Deputy Head of Profession for Statistics, Wales

Roger will describe the approach and challenges in measuring wellbeing of a nation, and how to measure and present change to the Government, Parliament and the public. Scotland's National Performance Framework is a statement of what the nation wants to achieve and how it will measure whether it is succeeding. Established in 2007, this was an innovative way of delivering Government by focussing on outcomes and wellbeing, and reporting progress based on official statistics. This session will cover how this approach is being aligned with the UN Sustainable Development Goals, how it forms an important part of Scotland's approach to Open Government, and drives the research needs of Government in Scotland.

7.3 Invited - Dependent Functional Data
Wednesday 6 September – 3.40pm-5pm

Boosting Generalized Additive Models for Location, Scale and Shape for Functional Data

Almond Stöcker, Sarah Brockhaus, Sophia Schaffer, Benedikt von Bronk, Madeleine Opitz, Sonja Greven
Department of Statistics, Ludwig-Maximilians-Universität, Munich

We extend Generalized Additive Models for Location, Scale, and Shape (GAMLSS) to regression with functional response. GAMLSS [Rigby&Stasinopolous, JRSS(C), 2005] are a flexible model class allowing for modelling multiple distributional parameters at once, such that for each parameter an individual General Additive Model predictor can be specified. By expanding this to functional regression we may for example simultaneously model the point-wise mean and variance of response curves over time in dependence of covariates. In addition, a variety of marginal probability distributions can be applied for response measurements, exceeding exponential families.

The model is fitted combining boosting based functional regression [Brockhaus et al, STAT COMPUT, 2017] and boosting based GAMLSS [Mayr et al, JRSS(C), 2012]. This provides inherent model selection and regularization. Moreover, simulation studies show that the overfitting control via curve-wise cross-validation is desirably sensitive to in-curve dependence.

We apply the functional GAMLSS to analyse bacterial interaction in E. coli and show how the consideration of variance structure fruitfully extends usual growth models. Employing historical function-on-function effects, we model the distribution of the growth curve of one bacterial strain depending on the complete progression of a competitive strain up to the current time point. By expending functional response regression beyond the mean, our approach offers new insights into underlying processes and helps overcome overly restrictive model assumptions.

7.3 Invited - Dependent Functional Data
Wednesday 6 September – 3.40pm-5pm

Analysis of Spatially Correlated Functional Data

Surajit Ray
University of Glasgow

Datasets observed over space and time have become increasingly important due to its many applications in different fields such as medicine, public health, biological sciences, environmental science and image data. Both spatiotemporal methods and functional data analysis techniques are used to model and analyse these types of data considering the spatial and temporal aspects. In this talk we will present an integral framework for modeling and analysing functional which are spatially correlated. In particular we wish to integrate existing approaches and identify gaps for analyzing a wide variety of spatially correlated functional data and provide the practitioner with objective choices to identify the best method to analyze their data.

7.3 Invited - Dependent Functional Data
Wednesday 6 September – 3.40pm-5pm

An Object-Oriented approach to Kriging and Uncertainty Quantification for Functional Compositional Data with Spatial Dependence

Alessandra Menafoglio, Piercesare Secchi, Alberto Guadagnini
Politecnico di Milano

Modern field studies yield diverse types of observations, in the form of highly heterogeneous and high-dimensional data. In this context, sedimentological and/or geochemical observations are becoming increasingly available in the form of functional (e.g., curves, surfaces or images) or distributional data (e.g., cumulative distribution or probability density functions). Menafoglio et al. (2013) propose to treat all these diverse types of information within a unifying framework. They consider the available observations as “object” data so that the building-block of the geostatistical analysis is the entire object, rather than a limited number of selected features of the data.

Motivated by the operational challenge of providing predictions and uncertainty assessment of the three-dimensional spatial distribution of particle-size curves (PSCs) within a field scale heterogeneous alluvial aquifer, we place key emphasis here to the problem of prediction (i.e., Kriging) and uncertainty assessment (via stochastic simulation) of distributional data. We interpret the latter as points within the Hilbert space of functional compositions, endowed with the Aitchison geometry. This, in turn, enables one to: (i) define an appropriate notion of spatial dependence and accordingly perform linear predictions through an (Object-Oriented) Universal Kriging approach; and (ii) effectively reduce the data dimensionality for the purpose of stochastic simulation.

We test these methods by way of a real field application relying on a set of particle-size curves collected through sieve-analysis at a well-documented test site. We remark that the methodology is entirely general and open to a broad range of environmental and industrial applications. It also opens new perspectives in the geostatistical analysis of data belonging to manifolds (Pigoli et al., 2016).

7.4 Invited - 75th anniversary of the Beveridge Report
Wednesday 6 September – 3.40pm-5pm

Beveridge's relevance today

Frank Popham
University of Glasgow

Beveridge‘s work, particularly his famous 1942 report, was crucial in shaping the post-war welfare state. It aimed to tackle the five giants of want, disease, ignorance, squalor and idleness. In this talk I will briefly outline how Beveridge’s work influenced the post-war welfare state. I will then focus on three present day concerns of income inequality, full employment and universality and how Beveridge’s legacy can inform debates on these topics.

7.4 Invited - 75th anniversary of the Beveridge Report
Wednesday 6 September – 3.40pm-5pm

Welfare, poverty and benefit fraud: long term patterns and recent developments using data from the British Social Attitudes survey

Eleanor Attar Taylor
National Centre for Social Research

Since 1983 the British Social Attitudes survey (BSA) has been annually measuring and tracking changes in people's social, political and moral attitudes. Drawing on 34 years of BSA data this talk will present trends in attitudes to the welfare state, benefit recipients and poverty; exploring how views of social welfare have been affected by economic and societal shifts as well as policy changes over the past three decades. It will also delve into new data on the extent to which benefit recipients are viewed as ‘deserving’, the perceived prevalence of benefit fraud in Britain and how views of benefit fraud compare with that of tax evasion.

7.4 Invited - 75th anniversary of the Beveridge Report
Wednesday 6 September – 3.40pm-5pm

The influence of welfare reform in the United States on the UK welfare state

Marcia Gibson
MRC/CSO Social and Public Health Sciences Unit

Reforms to social security benefits in the UK were influenced by the experience of welfare reform in the United States. Beginning in the early 1990s, welfare benefits for lone parents in the USA were dramatically scaled back. Employment increased rapidly, and welfare rolls dropped substantially. Welfare reform was hailed as a great success, and this strongly influenced reforms to the UK social security system. Prominent amongst these is the introduction of mandatory work requirements for lone parents, with removal of benefits for failure to comply with requirements.

This talk considers how a simplistic interpretation of statistics relating to the effects of major policy change in the United States was used to justify the erosion of rights-based entitlement to social security, previously a key plank of the UK welfare state. It also discusses the insights that can be provided by different approaches to evaluating the impacts of population-level policy interventions, and the ways in which a more complete understanding of contextual factors alters our interpretation of the effects of welfare reform in the US. The talk draws on a systematic review of welfare reform studies, a natural experiment assessing the impact of recent social security reforms in the UK, and contextual data which shed light on the longer terms effects of welfare reform in the US.

7.5 Invited - Recent advances in Statistical Signal Processing
Wednesday 6 September – 3.40pm-5pm

A framework for stochastic process modelling of complex-valued signals

Adam Sykulski
Lancaster University

In many applications of signal processing, bivariate signals are represented as complex-valued signals. This representation is useful for separating signals that are circular vs noncircular (sometimes referred to as proper vs improper). In this talk, we present a framework for the parametric modelling of such signals using stochastic processes. We apply our framework to two applications. The first uses a novel widely-linear autoregressive process to model noncircular seismic signals. The second uses a novel anisotropic Matern process to model time series obtained from particle trajectories seeded in fluid dynamic models of turbulence.

7.5 Invited - Recent advances in Statistical Signal Processing
Wednesday 6 September – 3.40pm-5pm

A spatial statistics approach to assessing signal extraction quality and defining resolution in modern optical imaging

Ed Cohen
Imperial College London

For centuries, optical imaging has been the primary tool by which humans have observed objects at microscopic and astronomic scales. Resolution is a fundamental property of an optical system, yet it remains difficult to define and apply. The classical notions of resolution such as Abbe's and Rayleigh's criteria had a major influence on the development of optical methodologies. However, they were developed in the context of observations with the human eye and are not appropriate for modern optical imaging methods that cannot be separated from the sophisticated signal and image processing algorithms that are used to analyse the acquired data. Spatial point processes have been extensively used in a range scientific disciplines to model spatial data with imaged objects being frequently analysed using spatial statistics approaches such as Ripley's K-function and the pair correlation function. We show that insufficient resolution can have a significant impact on the recovery and analysis of spatial point patterns. We quantify this effect and in doing so we use spatial statistical theory to provide new measures of optical resolution as well as introducing the concept of algorithmic resolution. We motivate this work with the application of super-resolution microscopy which has allowed biological samples to be imaged in unprecedented detail. But just how super is super-resolution?

7.6 Invited - Communicating statistics via social media, technology and blended learning
Wednesday 6 September – 3.40pm-5pm

Teaching statistics for OR and Finance to Post-graduate using VLE such as Blackboard

Vesna Perisic
University of Southampton

The talk will focus on teaching a university statistics course designed for master students in OR and OR and Finance and will demonstrate innovative ways to teach statistics using technology.

Both of our MSc in Operational Research programmes, MSc OR and MSc OR & Finances, include a statistics module. Given academic and other diversities of the cohorts which currently have up to 30 students, teaching of this module is not without challenges. Though some of the students have done mathematics undergraduate degrees, many come with no statistical knowledge at all.

The module aims to provide the students with sufficient knowledge of statistics to enable them to carry out simple statistical procedures when they arise, both in their summer projects and their later working life. Therefore, we are more focused on teaching practical applications of the statistical techniques emphasizing interpretation and communication of ideas to both specialists as well as non-specialists. MINITAB software package is an integrated part of the module used to support the lectures. A few laboratory sessions are offered to make the students familiar with the software. Our teaching approach is described in [1].

The session will give insight into how we successfully use a Virtual Learning Environment (VLE) such as Blackboard, in support of our teaching and assessment of this MSc module. Furthermore, Blackboard is not only used to deposit the teaching material, submit an assessment or communicate feedback individually to the students, but also provides an easy way to our external examiners for their external scrutiny of the module.

Different dimensions of utilizing the VLE within this statistics module will be illustrated on the examples from our practice gained from teaching this module.

Reference:

1. Christine S.M. Currie; Vesna Perisic : Minitab: the natural choice for non-specialist statisticians? , MSOR Connections Vol 9 No 3 August-October 2009

7.6 Invited - Communicating statistics via social media, technology and blended learning
Wednesday 6 September – 3.40pm-5pm

Teaching statistics to medical researchers and enabling them to teach us

Jamie Sergeant
University of Manchester

In this talk I will discuss teaching statistics to clinicians and other non-statisticians who need to access or undertake medical research. I will consider different possibilities for reaching such an audience through channels specific to their own field, for example conferences and journals. Although spreading the gospel of good statistical practice sounds altruistic, statisticians may be motivated to take part in such activities by hoping to enable better-informed peer review of their own applied work. I will also describe how clinicians and other specialists might be enabled to teach their subject-specific knowledge to statisticians and other non-specialists working in applied fields, where typically little or no training may be available. An example project is Rheum 101, a series of short accessible talks by clinicians on clinical rheumatology topics, broadcast live and then available as a resource on YouTube.

7.6 Invited - Communicating statistics via social media, technology and blended learning
Wednesday 6 September – 3.40pm-5pm

Reducing statistics anxiety via social media

Meena Kotecha, Jamie Sergeant, Vesna Perisic
The London School of Economics and Political Science

This interactive presentation will demonstrate an innovative approach to reducing statistics anxiety.

Students can find it difficult to engage with statistics due to non-cognitive factors, such as negative attitudes and/or pre-conceived notions about statistics, which can obstruct their learning.

Statistics anxiety has an adverse impact on students' confidence and their academic performance. This is a serious issue which can have an impact on students’ future career choice.

Facebook is used in this research informed teaching practice developed over a longitudinal study aimed at understanding and reducing statistics anxiety, promoting inclusive practice and enhancing student interaction as well as their learning experience.

The rationale behind this is to use a platform that is familiar to students and associated with fun, in order to create a relaxed learning climate; enhance student engagement and create interest in statistics.

Academics from all related disciplines should be able to apply the proposed techniques to delivering any quantitative courses. Furthermore, this should be of interest to statistics education researchers and all interested in the theme.

7.7 Invited - Recommender systems
Wednesday 6 September – 3.40pm-5pm

Diversity as a Response to User Preference Uncertainty

James Edwards, David Leslie
Lancaster University

This talk addresses the problem of recommending a set of objects (e.g. purchases, videos or news stories) to a user. We wish to choose objects that are most appropriate for the user but selecting on this criterion alone can lead to a set that is insufficiently diverse which results in redundancy in the set.

The usual approach to this problem is to explicitly trade off appropriateness and diversity as two separate objectives. In contrast we build on an approach from information retrieval, in which diversity emerges as an extrinsic need. By taking into account uncertainty about the preferences of the user, diversity naturally arises from the need to 'cover the bases' when maximising the overall probability that the user chooses an object. The level of diversity is thus automatically tailored appropriately for the level of uncertainty in the user's preferences. In addition, we show that the choice of model for user behaviour strongly affects the diversity of the optimal set.

7.7 Invited - Recommender systems
Wednesday 6 September – 3.40pm-5pm

From recommender systems to algorithmic decision-making

Sofia Olhede, Russell Rodrigues, Patrick Wolfe
UCL

Recommender systems are a linchpin of modern online retail systems, and have garnered significant attention, including the Netflix prize of 2006. They are generally evaluated by predictive performance, and generate considerable success for the retail sector. Online retail recommendations are of less ethical sensitivity, but as algorithms are increasingly embedded in a broader range of societal applications, the repercussions of their use may become considerably more consequential. I will discuss some recent developments in data governance and algorithmic decision-making in areas beyond retail.

7.7 Invited - Recommender systems
Wednesday 6 September – 3.40pm-5pm

Bayesian preference learning with the Mallows rank model for explicit and implicit data

Arnoldo Frigessi
University of Oslo

The Mallows rank model can be used to learn individual rankings of items, based on incomplete preferences expressed by users in explicit form (ratings, ranking, comparisons, …). We recently developed a computationally feasible approach to Bayesian inference for Mallows models that works with any right-invariant distance. I will summarise our method first. Then I will move to implicit data, as most preference data today are in implicit form, for example clicks on items expressing likes. We extend our Mallows approach to implicit data and compare it to some versions of collaborative filtering, which is the most popular approach, in terms of precision of the recommendation and computational scalability when the number of assessors and/or items is very large. This is joint work with Valeria Vitelli, Marta Crispino, Øystein Sørensen, Elja Arjas, Sylvia Qinghua Liu.

7.8 Invited - Data-Driven DoE - Using penalised regression to unify the analysis and enhance the practice of industrial experimentation
Wednesday 6 September – 3.40pm-5pm

Data-Driven DOE: A Case Study

Volker Kraft
SAS Institute / JMP Division

Using statistically designed experiments (DOE) is the best approach to learning from data, since it has the potential to be both efficient and effective. However, realizing this potential relies on clearly articulating what you already know, and what you wish to learn. If they exist, observational data can help to establish what is ‘known’, but handling such data appropriately can be difficult. However, if these difficulties can be addressed, one can exploit more coherent cycles of learning that leverage data to the full - ‘Data-driven DOE’ for short.

The case study will show an example how Data-driven DOE was applied to improve a manufacturing process for liquid crystals displays (LCDs): Historically, the pigment milling step had caused many problems, with long mill times that were also extremely variable. Even though only a small and messy data set of historical process measurements was available, a significant improvement could be effected in just one cycle of learning - Situation appraisal, designed experiment, modeling and optimization. A live demonstration will show the challenges and solutions during all these steps.

7.8 Invited - Data-Driven DoE - Using penalised regression to unify the analysis and enhance the practice of industrial experimentation
Wednesday 6 September – 3.40pm-5pm

Designed experiments for interaction screening

David Woods
University of Southampton

Product and process improvement can involve a large number of factors which must be varied simultaneously. Understanding how factors interact is a key step in identifying those factors that have a substantial impact on the response. We provide an assessment and comparison of screening strategies for interactions using carefully designed experiments and a variety of data analysis methods including shrinkage regression and Bayesian methods. Insights on using the strategies are provided through a variety of simulation scenarios and open issues are discussed.

8.1 Contributed - Medical statistics: Epidemiology and infectious disease modelling
Thursday 7 September – 9am-10am

Assessing the impact of a temporary class drug order on ethylphenidate-related infections among people who inject drugs: an interrupted time-series analysis

Alan Yeung, Andrew McAuley, Amanda Weir
Health Protection Scotland

Background: The increasing availability of novel psychoactive substances (NPS) is creating considerable challenges for public health. Legislation has been used in response to rises in NPS use and related harms, however few studies have been conducted evaluating their effectiveness. In April 2015, the UK government enacted a temporary class drug order (TCDO) on ethylphenidate in response to reported harms associated with its use, in particular an outbreak of infections among people who inject drugs (PWID) in Scotland. We assess the effect that the TCDO had on reducing the most common infections identified during the outbreak due to Streptococcus pyogenes (S. pyogenes) and Staphylococcus aureus (S. aureus).

Methods: The outbreak was split into a pre-intervention period (35 weeks) and a post-intervention period (26 weeks) based around the date of the TCDO. Segmented Poisson regression was used to compare trends in weekly counts of infections. Self-reported NPS use was included in the model to allow relative risks (RRs) of infection to be estimated for NPS users compared with those with no self-reported NPS use.

Results: There were 251 S. pyogenes and/or S. aureus infections recorded among 211 PWID between February 2014 to December 2015. Significant trend changes in weekly S. pyogenes and/or S. aureus infections following the TCDO were found (RR 0.87, 95% CI 0.83–0.91). NPS users were at higher risk of acquiring these infections (RR 1.92, 95% CI 1.46–2.53), particularly when comparing NPS use for S. pyogenes emm76.0 against S. pyogenes (emm types other than emm76.0) (RR 3.60, 95% CI 1.43–9.04).

Conclusions: The ethylphenidate TCDO was effective in reducing infections among PWID during an outbreak situation in Lothian, Scotland. Legislative interventions aimed at decreasing accessibility and availability of particular substances can play an important role in the public health response to disease outbreaks linked to use of NPS.

8.1 Contributed - Medical statistics: Epidemiology and infectious disease modelling
Thursday 7 September – 9am-10am

Comparison of statistical algorithms for the detection of outbreaks in syndromic surveillance systems

Roger Morbey
Public Health England

Syndromic surveillance involves monitoring big health datasets to provide early warning of threats to public health. Public health authorities use statistical detection algorithms to mine these datasets for aberrations that are indicative of emerging threats. The algorithm currently in use at Public Health England (PHE) for syndromic surveillance is the 'rising activity, multi-level mixed effects, indicator emphasis' (RAMMIE) method (Morbey *et al*, 2015), which fits a mixed model to counts of syndromes on a daily basis. The aim of this study is to investigate whether alternative statistical approaches can improve detection at different levels or whether RAMMIE is adequate for syndromic surveillance. For this purpose, we compare RAMMIE to the improved quasi-Poisson regression-based approach (Noufaily *et al*, 2013), currently implemented at PHE for weekly infectious disease laboratory surveillance, and to the cumulative sum control charts (CUSUMs) method (Rossi et al, 1999), which is used for syndromic surveillance aberration detection in many other countries. We model syndromic datasets, capturing real data aspects such as long-term trends, seasonality, bank holidays, and day-of-the-week effects, with or without added outbreaks. Then, we compute the sensitivity and specificity to compare how well each of the algorithms detects outbreaks to provide recommendations for the most suitable statistical methods to use during different public health scenarios.

Morbey, R. A., Elliot, A. J., Charlett, A., Verlander, A. Q, Andrews, N. and Smith, G. (2013). The application of a novel 'rising activity, multi-level mixed effects, indicator emphasis' (RAMMIE) method for syndromic surveillance in England, *Bioinformatics*, 31(22), 3660-3665.

Noufaily, A., Enki, D. G., Farrington, C. P., Garthwaite, P., Andrews, N. and Charlett, A. (2013). An Improved Algorithm for Outbreak Detection in Multiple Surveillance Systems. *Statistics in Medicine*, 32(7), 1206-1222.

Rossi, G, Lampugnani, L, Marchi, M. (1999), An approximate CUSUM procedure for surveillance of health events. *Statistics in Medicine*, 18, 2111–2122

8.1 Contributed - Medical statistics: Epidemiology and infectious disease modelling
Thursday 7 September – 9am-10am

Estimating the force of infection of blood-borne viruses in people who inject drugs: key risk factors and frailty

Ross Harris, Daniela De Angelis, Vivian Hope, Paddy Farrington
Public Health England

Serological survey data on long-lasting antibodies may be used to estimate age-specific force of infection (FOI) and individual frailty via bivariate associations between infections with shared transmission routes. We used such models to examine risks of hepatitis C (HCV) and hepatitis B infection in people who inject drugs. Monitoring and understanding risks of infection is important for planning harm reduction interventions.

Unlinked anonymous monitoring data were available from 2000 to 2014 and include antibody tests and questions on injecting duration (analogous to age) and a number of risk factors. Previous modelling was extended to include covariates and estimate hazard ratios (HR) associated with different risk factors and demographics; and inverse Gaussian and time-varying frailty distributions to investigate individual frailty and how this evolves over time after first injection.

Results indicated a substantial excess risk in the first year of injecting (HR=3.21 for HCV) and broadly constant thereafter (5% per year). Risk was higher in females (HR=1.38), those ever imprisoned (HR=1.48) and those reporting needle sharing (HR=1.56). Moderate heterogeneity remained after adjusting for covariates, equating to a HR of 3.9 between 75th and 25th percentiles of risk. Relative frailty variance was estimated to decline by around a half after 10 years of injecting. Risk in males was more variable than that of females, particularly in recent initiates to injecting.

This analysis highlights an excess risk in recent initiates, females and vulnerable groups. However, there is marked residual frailty, especially early in injecting career. Early engagement of recent initiates and those at high risk is important for harm reduction, although substantial residual frailty means that identifying high-risk individuals may be difficult. Work is underway to extend the bivariate frailty model to a trivariate model that also includes HIV, to better understand components of individual frailty that may also include sexual transmission.

8.2 Contributed - Official statistics and public policy: Money matters
Thursday 7 September – 9am-10am

Economic statistics classifications; are they a road block to infrastructure investment?

Derek Bird, David Beckett, David Bailey
Office for National Statistics

With government deficit and debt becoming ever more a political issue, economic statistics classifications have never been so high profile or as consequential to the public. Governments across the UK have committed to reversing the classification of housing associations from the public sector, which has added almost £100bn of debt to the government books.

Future capital investment in schools, hospitals and transport infrastructure is contingent upon projects' economic statistics classification and the fiscal rules which are aligned with the statistical classification; while every University in the UK is under review to determine the institutional sector in which they should be placed. The paper will set out the process and primary considerations that underpin the statistical decisions that see programmes appearing on or off the government's balance sheet.

In this session Derek Bird, the head of the Office for National Statistics’ Public Sector Division, will share his experience of how the ONS, Treasury, Bank of England and Downing Street work together on these issues, and what recent events mean for public policy.

8.2 Contributed - Official statistics and public policy: Money matters
Thursday 7 September – 9am-10am

Small Area Estimation of Fuel Poverty in England

Katie Allison
Department for Business, Energy & Industrial Strategy

This presentation investigates improving methods for estimating sub-regional fuel poverty levels.

To ensure official statistics remain relevant, robust, reliable and transparent, the sub-regional fuel poverty statistics, published by the Department for Business, Energy and Industrial Strategy, were reviewed and an alternative method devised and compared to the previous estimates.

This presentation takes a look into the development of an alternative methodology (results pending) with the focus on the following:

- Developing a multi-level logistic regression model with a binary survey response variable sourced from the English Housing Survey and with admin data as the dependent variables.
- The model selection used in order to determine the best fitting model – taking into consideration the reliability and timeliness of the admin data.
- Producing corresponding uncertainty estimates.
- Consideration of alternative methods including machine learning and Bayesian methods.

Developing an alternative methodology has allowed us to move away from being reliant on modelled Experian data, produce corresponding uncertainty estimates (which were previously not produced) while improving and utilising modelling expertise within the GSS.

8.2 Contributed - Official statistics and public policy: Money matters
Thursday 7 September – 9am-10am

#FlowofFunds - understanding how money moves

Richard Campbell
Office for National Statistics

In 2008, the world changed forever. The financial crisis shook our economy and unveiled worrying holes in our understanding of how it works.

The objective of the Enhanced Financial Accounts (also known as the ‘Flow of Funds’) initiative is to fill these holes through innovative use of public and private sector data. The initiative will develop policy maker and public understanding of how money moves around the economy – who owes who and where the greatest risks are lurking.

This session will cover a number of aspects of the work. Firstly, it will provide context for the work – helping the audience understand why it is important and why ONS and the Bank of England are working on it. Secondly, it will explain in detail the questions the need answering and the approach being taken to do so. Thirdly, it will explore how ONS is working in partnership with the private sector to gain unparalleled access to new data sources. Finally, it will present early results from the initiative and explain what they mean for the future of the initiative.

The success of this initiative relies on successful collaboration with as broad a range of interested parties and experts as possible. To help enable this, the session will be interactive – giving the audience the opportunity to share their expertise and policy needs with the presenter.

8.3 Contributed - Environmental/Spatial statistics: Environmental behaviour and forecasting
Thursday 7 September – 9am-10am

Probabilistic Energy Forecasting and Applications

Jethro Browell, David McMillan, Matthew Revie, Ciaran Gilbert
University of Strathclyde

In the UK, and internationally, the way energy is produced and consumed is changing dramatically in order to mitigate the negative impacts of climate change. The growth in weather dependent electricity generation combined with the increase in the dependence of electricity demand on weather due to the electrification of heat and transport is increasing the sector's reliance on weather forecasting. This increasing reliance on forecasts brings increased uncertainty of future supply and demand on operational timescales. As a result, operational practices must adapt to maintain high standards of reliability and economic efficiency of both power systems and energy markets. Part of that adaptation is the transition from single-value "point" forecasts to probabilistic forecasts, which quantify forecast uncertainty and are optimal inputs to most decision-making problems. However, the complexity of probabilistic information is a barrier to the adoption of probabilistic forecasts within the energy industry and across society in general. Here, examples of probabilistic energy forecasting and applications to specific decision-making problems will be presented.

Maintenance for wind turbines may require use of cranes to lift large components or, in the case of offshore wind, personnel to transfer from boat to turbine. These operations are subject to safety limits on wind speed and wave height, respectively, which must be forecast. Scheduled operations are therefore subject to forecast error. Providing planners with actionable probabilistic information can support them in optimising schedules and managing risk.

Modern power systems are increasingly weather dependent and exhibit significant spatio-temporal and cross-resources structure. Errors in weather forecasts may be realised as errors in wind, solar and demand forecasts which may compound or oppose each other depending on the situation. Capturing this structure is required to efficiently manage power system security and support energy market participants. An example of large-scale energy forecasting will be presented here.

8.3 Contributed - Environmental/Spatial statistics: Environmental behaviour and forecasting
Thursday 7 September – 9am-10am

A non-homogenous Poisson model with spatial anisotropy and change-points: an application to ozone data

Eliane Rodrigues, Geoff Nicholls, Mario Tarumoto, Dani Gamerman, Guadalupe Tzintzun
Universidad Nacional Autonoma de Mexico (UNAM)

We consider a non-homogeneous Poisson model to estimate the probability that an environmental threshold is exceeded a given number of times in a time interval of interest. Besides depending on time, the rate function of this Poisson model will also depend on some parameters. We also allow the presence of change-points in the model. An anisotropic spatial component is imposed on the vector of the parameters of this rate function as well as on the vector of possible change-points. The parameters of the rate function and the parameters of the spatial model as well as the location of the change-points will be estimated using the Bayesian point of view via a Metropolis-Hastings algorithm within the Gibbs sampling. The model is applied to ozone data obtained from ten stations which are part of the monitoring network of Mexico City, Mexico. Each station will have its own Poisson model assigned to it and their interaction will be via the spatial model applied jointly to all of them in the parameters of the rate function and change-points. In the application we consider the maximum daily ozone measurements obtained from 01 January 1990 to 31 December 2010. Results suggest that two change-points are necessary to have a good fit of the accumulated observed and estimated mean functions in each station. They also indicate that the behaviour of the rate function is decreasing with smaller rates as we go towards the end of the observational period. This might be an indication that the several measures taken by the environmental authorities during that time, in order to reduce population exposure to high levels of ozone, have had a positive effect. This is a joint work with Geoff Nicholls, Mario H. Tarumoto, Dani Gamerman, and Guadalupe Tzintzun.

8.3 Contributed - Environmental/Spatial statistics: Environmental behaviour and forecasting
Thursday 7 September – 9am-10am

Cross-National Comparisons on Pro-Environmental Behaviours – A Multilevel Modelling Approach

Vivian Hiu Tung So, Gabriele B Durrant, Peter WF Smith
University of Southampton

This study investigates the cross-national differences in pro-environmental behaviours by identifying how individual and country-level factors influence individuals’ behaviours using a multilevel modelling framework. Moreover, it also seeks to explore how personal environmental attitudes have different degrees of effect on how people behave across nations using random slope models. Analysis is conducted on the 2010 Environmental module of the International Social Survey Programme (ISSP), a cross-national survey that mainly deal with environmental behaviours and attitudes towards environmental related issues. General pro-environmental behaviour, as well as home-, purchase-, transport- and recycle-related environmental behaviours are considered. Since different types of environmental behaviours are not necessarily correlated, decomposing the unidimensional measure of general pro-environmental behaviour into multidimensional measures allows us to identify the underlying trends that may have been hidden during aggregation. Preliminary results show that both individual- and national-level variables have substantial effects in explaining different types of pro-environmental behaviours after controlling for sociodemographic factors and mode of interview. Moreover, the inclusion of the random slope in the final models also provides evidence that the effects of individuals’ environmental attitudes on their environmental behaviours vary significantly across countries.

8.4 Contributed - Applications of Statistics
Thursday 7 September – 9am-10am

Hidden Markov Models for Monitoring Circadian Rhythmicity in Rest-Activity Data

Qi Huang, Barbel Finkenstadt Rand, Francis Levi, Dwayne Cohen, Sandra Komarzynski, Innominato Pasquale, Xiaomei Li
University of Warwick

Wearable computing devices allow collection of real-time densely sampled information on movement enabling researchers and medical experts to obtain objective and non-obtrusive records of actual activity of a subject in the real world over many days. Our interest here is motivated by the use of activity data for evaluating and monitoring the endogenous circadian rhythmicity of subjects for research in chronobiology and chronotherapeutic healthcare. In order to translate the information from such high-volume and complex data arising we propose the use of a hidden Markov modelling approach which (a) naturally captures the notable square wave form with heterogeneous variances over the circadian cycle of human activity, (b) solves the problem of thresholding activity into different states in a probabilistic way respecting time dependence between successive observations and taking into account any additional information such as covariate measurements, circadian timing of the activity and other relevant information, and (c) delivers parameter estimates that are interpretable and important to circadian research.

8.4 Contributed - Applications of Statistics
Thursday 7 September – 9am-10am

Statistical Signal Processing: Applications in image denoising

Varuna De Silva
Loughborough University

With the emergence of the widespread use of smartphones, capturing images and videos has become ubiquitous in our daily lives. Advancement of sensor technology means now we are capable of capturing high resolution images at a very high quality. The image signal processing pipeline of an imaging device plays a crucial role in delivering a high quality image/video frame through processing and cleaning the raw sensor data that is captured by the sensor utilizing image processing algorithms. Image denoising is the process which tries to reduce the artifacts of the data acquisition stage (i.e. noise) due to imperfections in the sensing technology. Noise can be characterized by various distributions, and often represented as a Gaussian distribution.

Image denoising is a classical inverse problem. Over the last decade researchers have exploited various statistical models of natural images to develop effective methodologies to overcome sensor noise. In this paper we will provide an overview of the effective use of statistics in this process. The paper will present various noise distributions, natural image statistical models and state of the art methods to overcome noise. We will particularly focus on Wiener filtering (optimal in a mean squared error sense) and Bayesian methods for noise reduction. Finally we will describe system level challenges to implement statistical methods on consumer electronic devices.

8.4 Contributed - Applications of Statistics
Thursday 7 September – 9am-10am

Statistical modelling of cell movement data using the unscented Kalman filter

Diana Giurghita, Dirk Husmeier
University of Glasgow

The unscented Kalman filter (UKF) has been extensively used in the simultaneous estimation of states and parameters in nonlinear state space models. At the core of the UKF is the unscented transform, which is used to approximate Gaussian distributions in nonlinear settings, using a deterministic set of sigma points. Recent studies identify an area for improvement for this algorithm by optimising the parameters that determine the placement of the sigma points.

In this talk, we compare the performance of default and optimized parameters, which we identify using Bayesian optimisation. We also demonstrate an application of the improved UKF in the context of cell movement obtained from high resolution microscopy data. Our study in collective cell movement is motivated by the widely observed movement patterns in many important biological processes, such as wound healing, tissue development and cancer cell migration.

8.5 Contributed - Methods and theory: Longitudinal & Causal Models
Thursday 7 September – 9am-10am

Singular Linear Models for Longitudinal Data

Daniel Farewell, Chao Huang
Cardiff University

We consider the analysis of longitudinal data in settings where the number and timing of measurements may be informative about their recorded values. Heuristically, bias arises because subjects with more observations contribute disproportionately to estimation, but this bias can be reduced through a working assumption that the within-subject covariance matrix is singular. Singular linear models (SLIMs) arise as limiting solutions to estimating equations in which the working covariance tends to such a reduced-rank matrix. Because of their relative computational simplicity and the paucity of assumptions required, SLIMs offer an attractive alternative to joint modelling and inverse probability weighting. We describe the ability of SLIMs to accommodate dynamic dependence between timings and measurements, different covariance structures, and multivariate longitudinal data, illustrating their use in standard software.

8.5 Contributed - Methods and theory: Longitudinal & Causal Models
Thursday 7 September – 9am-10am

A testing strategy for choosing between estimation methods for random effects models

Daniela Bond-Smith
Centre for Genetic Origins of Health and Disease, University of Western Australia

Standard (restricted) maximum likelihood (ML/REML) estimation methods for random effects models tend to underestimate uncertainty, producing unduly narrow confidence intervals. Alternative methods have been developed to address this weakness, but they tend to generate less accurate point estimates. Hence, users face a trade-off between accuracy and prudence. There is little guidance for applied users on how to decide whether estimation results are reliable and when a change in estimation approach is warranted. In particular for moderately sized samples, established knowledge about estimator performance for very small or large samples is ambiguous and less tangible factors than sample size are critical. This paper proposes a simple, intuitive testing strategy to support users' assessment of the plausibility of their estimates. The suggested approach capitalises on the observation that differences in estimate distributions generated by ML/REML estimators and those from a weakly-informative Bayesian method that avoids boundary estimates for the heterogeneity parameter are indicative of data characteristics where standard estimators have significant weaknesses in generating reliable estimates of uncertainty. Simulation results in the paper confirm the intuition that comparing the probability distributions of estimates performs well as a proxy for the ideal, but theoretical case where the true parameters are known. The results also demonstrate that the two estimation strategies are complementary for different data characteristics and the proposed test identifies a suitable switching point according to key objectives, such as ensuring adequate coverage and avoiding underestimation of uncertainty. A simple flow chart for applied users is provided that indicates when the proposed test is particularly useful and an R package is currently being written. The key intent of the paper is to support improved estimation strategy choices in an applied setting on the basis of the available theory and to help make important statistical considerations more easily digestible for the applied user.

8.5 Contributed - Methods and theory: Longitudinal & Causal Models
Thursday 7 September – 9am-10am

On Causal Estimation using U Statistics

Lu Mao
University of Wisconsin-Madison, USA

We introduce a general class of causal estimands which extends the familiar notion of average treatment effect. This new class of estimands is defined by a metric function, pre-specified to quantify the relative favourability of one outcome over another, averaged over the marginal distributions of the two potential outcomes. In the hypothetical situation with fully observed counterfactual data, natural estimators arise in the form of U statistics. Under standard assumptions on the observed data with pre-treatment confounders, the inverse propensity score weighting technique is applied to the kernel of the U statistic to achieve unbiased estimation. In addition, a class of locally efficient and doubly robust estimators are derived using Hoeffding's decomposition theorem for U statistics combined with semiparametric efficiency theory. General approaches to construction of the locally efficient U statistic estimators are proposed and their robust variance estimators derived. The usefulness of our theory is illustrated by two examples, one for causal estimation with ordinal outcomes, and the other for causal tests that are robust to outliers.

8.6 Contributed – Medical statistics: Bioinformatics
Thursday 7 September – 9am-10am

A Random Effects Model for the Identification of Differential Splicing (REIDS) Using Exon and HTA Arrays

Marijke Van Moerbeke, Adetay Kasim, Ziv Shkedy
Hasselt University

Alternative splicing is a mechanism in which a single gene gives rise to multiple transcript isoforms. It was first believed to be an uncommon phenomenon. Recently, however, high-throughput sequencing technology has found indications that it affects more than 90% of human genes (Wang et al., 2008 ; Pan et al., 2008).

It has been shown that about 15% of the single nucleotide mutations in the exon recognition sites are associated with a human genetic disease. Many straying splice variants have been linked to cancers like mammary tumour genesis and ovarian cancer (Fan et al., 2006). Therefore, understanding the mechanism of alternative splicing and identifying the difference between splicing events in diseased and normal tissues is key in cancer research (Bisognin et al., 2014). In this study, we propose a hierarchical Bayesian model for alternative splicing detection. The Random Effects for Identification of Differential Splicing (REIDS) model identifies a splicing event based on two sets of scores; the array score which is used to identify samples that express an alternatively spliced exon and the exon score which is used to prioritize spliced probe sets.

Both array and exon scores are calculated using the parameter estimates for the variance components of the hierarchical Bayesian model. Furthermore, the Bayesian REIDS model allows to summarize the exon level data into gene level data in order to perform inference between the experimental conditions. The proposed model was applied to a variety of data sets which include exon arrays and HTA arrays. In the latter, the potential of exon-exon junctions was investigated. We concluded that exons which are supported by their junctions are more reliable alternative splicing candidiates. An R package (REIDS) in which the REIDS model is implemented is publically available in R-Forge.

8.6 Contributed – Medical statistics: Bioinformatics
Thursday 7 September – 9am-10am

Learning from our mistakes: Using sequencing errors to infer precision in mutation validation experiments.

Daniel Andrews, Inma Spiteri, Simon Tavaré, Andy Lynch
Illumina

One purpose of sequencing in cancer research is to identify somatic point mutations in the genome. Whole-genome sequencing experiments are seldom conducted to a depth that would allow us to infer precisely the proportion of cancer cells that carry a particular mutation - knowledge that would allow us to say something about the clonal structure of the tumour. Consequently, high-depth targeted sequencing is often conducted to make this inference with adequate precision. Naive interpretation of such experiments can encourage us to believe that we can achieve arbitrary precision through increased depth of sequencing. This is however illusory, and identifying the true precision achieved can be problematic.

In this talk we set out a framework for estimating precision by modelling the errors that are introduced into the sequences at the PCR amplification step. We also determine the circumstances in which the approach will work, and apply the method to a purpose-generated data set. One of the characteristics of our experimental design is that, unusually, we can construct an orthogonal estimate of the precision, which we will see validates our new approach.

A final observation is that such experiments typically use a high-fidelity polymerase that makes few errors. In the context, there would be little cost of using a lower-fidelity polymerase in estimating the mean, but potentially great benefit to estimation of the precision. Thus we conclude by adopting the provocative position of recommending the utilisation of cheap error-prone polymerases in such experiments.

8.6 Contributed – Medical statistics: Bioinformatics
Thursday 7 September – 9am-10am

Identification of transcript regulatory patterns in cell differentiation

Arief Gusnanto, John Paul Gosling, Christopher Pope
University of Leeds

Haematopoiesis is a formation of mature blood cells from their precursor stem cells. In the process, a stem cell will experience changes in gene expression and other complex processes that will direct it to a specific mature cell type. As a stem cell matures, it undergoes changes in gene expression (transcripts) that limit the cell types that it can become and moves it closer to a specific cell type. Studying transcript regulatory patterns in cell differentiation is critical in understanding its complex nature of the formation and function of different cell types. This is done usually by measuring gene expression at different stages of the cell differentiation. However, if the gene expression data available are only from the mature cells, we have some challenges in identifying transcript regulatory patterns that govern the cell differentiation.

In this study, we propose to exploit the information of the lineage of cell differentiation in terms of correlation structure between cell types. We assume that two different cell types that are close in the lineage will exhibit many common genes that are co-expressed relative to those that are far in the lineage. Current analysis methods tend to ignore this correlation by testing for diffferential expression assuming some sort of independence between cell types. We employ a Bayesian approach to estimate the posterior distribution of the mean of expression in each cell type, by taking into account the cell formation path in the lineage. This enables us to infer genes that are specific in each cell type, indicating the genes are involved in directing the cell differentiation to that particular cell type. We illustrate the method using gene expression data from a study of haematopoiesis.

8.7 Contributed - Data science: Applications and Methodology
Thursday 7 September – 9am-10am

Predicting the Results of the 2016 US Presidential Election

Timothy Martyn Hill
LV

This is the latest in a series considering the effectiveness of metrics at elections. This covers the 2016 US Presidential Election and is the companion to two Significance articles published in 2016 on that election.

Metrics are created by others and used to predict the outcome the 2016 US Presidential Election. So the question arises: which ones are the best? To answer this question we list the common metrics (polls, odds, models), we note the different level of detail for each metric (two candidates, three candidates, four candidates), and the methods we will use to measure their accuracy. We then use the chosen methods to compare those metrics to the final results, both within classes (which poll the best poll, which model the best model...), between classes (which is better: poll, model, odds...), 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: Applications and Methodology
Thursday 7 September – 9am-10am

Methods for compiling price indices from web scraped clothing data

Chris Payne
Office for National Statistics

The measurement of clothing prices presents a unique set of problems. Fashion items are typically very fast moving, which makes it difficult to follow items through time. Indeed, a change to collection practices for clothing was a major contribution to the increasing ‘wedge’ between the Consumer Prices Index and the Retail Prices Index in 2010. Large alternative data sources may offer a potential solution; however, such sources also come with challenges: we often see much higher product turnover than in local price collection.

In this analysis we make use of a web scraped dataset of clothing prices from September 2013 to October 2015 to:

- understand the extent of product turnover (or ‘churn’) in clothing.
- investigate how effective new and emerging price index compilation methods are when applied to clothing data, with its unique pattern of churn.

We utilise time line plots to understand the pattern of churn in the data, and make use of a Cox regression model to capture an item’s potential hazard to churn. We consider three price index compilation methods:

1. an indirect index, which chain links monthly bilateral indices
2. an Intersection-GEKS (IntGEKS) index, which is a member of the family of GEKS methods that average over all possible chain links
3. the Fixed Effects index with a Window Splice (FEWS), which uses a fixed effects regression model with product dummy variables to estimate a fully quality Adjusted index

As expected, we identify a high rate of churn in the clothing data. The IntGEKS and FEWS methodologies behave badly, with item indices decreasing by as much as 80% over the period. By contrast, the indirect index has a more sensible upward trend; however, the magnitude of seasonal movements is rather large.

8.7 Contributed - Data science: Applications and Methodology
Thursday 7 September – 9am-10am

Thoughts, Ruminations, and Twitter-ready soundbites on Data Science, Big Data, and Social Science Research

Craig Hill
RTI International

Using "Big Data" and data science tools, techniques, and approaches to produce (or predict) estimates is often viewed as a threat to the livelihood of statisticians and survey researchers. Often, these "new" tools or approaches are viewed with suspicion, or as being rife with error. In this talk, the author reviews what we know about big data and data science, examines several of the myths and buzzwords being promulgated about big data and data science, and provides several use cases for the audience's consideration. We start with definitions of "big data" and "data science" and then begin to unpack these definitions. For example, we delve into the catchphrase, "data is new oil." What does that mean, exactly? What are the implications for statisticians and others in our industry? Then, we examine and discuss several other aspects of the "big data" approach to social science research. Finally, we present several scenarios (or use cases) in which we have used data science-based approaches to social problems/issues for which we would have used a survey research-based approach in the past.

Plenary 6 - Barnett Lecture
Thursday 7 September – 10.10am-11.10am

Are you sure we want to do this? Sea level adaptation decisions under uncertainty.

Peter Guttorp, Thordis Thorarinsdottir, Martin Drews, Karianne de Bruin
Norwegian Computing Center

Sea level rise has serious consequences for harbor infrastructure, storm drains and sewer systems, and many other issues. Adapting to sea level rise requires comparing different possible adaptation strategies, comparing the cost of different actions (including no action), and assessing where and at what point in time the chosen strategy should be implemented. All these decisions must be made under considerable uncertainty–in the amount of sea level rise, in the cost and prioritization of adaptation actions, and in the implications of no action. Here we develop two illustrative examples: for Bergen on Norway's west coast and for Esbjerg on the west coast of Denmark, to highlight how technical efforts to understand and quantify uncertainties in hydrologic projections can be coupled with concrete decision-problems framed by the needs of the end-users using statistical formulations. Different components of uncertainty are visualized. We demonstrate the value of uncertainties and show for example that failing to take uncertainty into account can result in the median projected damage costs being an order of magnitude too small.

9.1 Invited - Statistical Causal Inference
Thursday 7 September – 11.40am-1pm

Detecting and correcting for pleiotropic bias in Mendelian randomization using gene-covariate interactions

Jack Bowden
University of Bristol

Mendelian randomization has developed into an established method for estimating causal effects, largely as a consequence of the proliferation of genome-wide association studies. However, the use of genetic instruments remains controversial, as pleiotropic effects can introduce bias into causal effect estimates. Recent work has highlighted the potential use of gene-environment interactions in detecting and correcting for pleiotropic bias in Mendelian randomization analyses.

We introduce linear Slichter regression (LSR) as a statistical framework capable of identifying and correcting for pleiotropic bias, drawing upon recent developments in economics and epidemiology. If a gene interaction exists which induces variation in the association between a genetic instrument and exposure, it is possible to identify the degree of pleiotropy and provide a corrected causal effect estimate. The interpretation of this method is similar to summary Mendelian randomization approaches such as MR Egger regression. A particular advantage of LSR is the ability to assess pleiotropic effects using individual instruments.

We investigate the effect of BMI upon systolic blood pressure using data from UK Biobank and the GIANT consortium, finding evidence to suggest the effect of BMI upon systolic blood pressure may be underestimated using conventional MR approaches. We assess the performance of LSR to identify and correct for horizontal pleiotropy in a simulation setting, and highlight the utility of the approach as a sensitivity analysis in cases where the method's assumptions are violated.

9.1 Invited - Statistical Causal Inference
Thursday 7 September – 11.40am-1pm

Data-adaptive estimation of high-dimensional mediated effects

Rhian Daniel
Cardiff University

In many modern biomedical applications, interest lies in decomposing the effect of an exposure on an outcome into its effect via a large number of mediators. For example, when trying to understand the mechanism through which a particular genetic variant affects a cardiovascular outcome, one might attempt to decompose the total effect of the variant into individual path-specific effects through hundreds of blood metabolite measurements.

Such an endeavour poses grave methodological challenges. First, the mediators are likely to be highly-correlated according to an unknown causal structure, including unmeasured common causes of one mediator and another, and the causal ordering of the mediators (if indeed there is one) is likely unknown. Second, the identification of natural path-specific effects, the most common choice of mediation estimand, in such a setting would rely on a large number of so-called "cross-world independence" assumptions, which are impossible to justify. Third, if we were to use a parametric estimation approach, as is most commonly done in mediation analysis, then, as the number of mediators increases, so too does the extent to which our inferences rely on incorrectly-specified and arbitrarily-chosen parametric nuisance models.

We propose that the first two problems be overcome by focussing on effects that are very similar to the interventional multiple mediator effects introduced by Vansteelandt and Daniel (Epidemiology, 2017). These avoid most cross-world independence assumptions and additionally allow agnosticism regarding the causal structure of the mediators, permitting unobserved common causes of one mediator and another. We propose that the third problem be overcome by adopting a data-adaptive estimation strategy. In this talk, I will give an overview of the proposed approach, demonstrating some of its properties in simulation studies.

9.1 Invited - Statistical Causal Inference
Thursday 7 September – 11.40am-1pm

Causality in Statistics

Robin Evans
University of Oxford

We provide a brief introduction to Statistical ideas about Causality and Causal Inference. There are several major schools of thought, each with differing perspectives on what causality means and how it should be distinguished from association. Particular points of controversy include: is it necessary to be able to design an experiment (at least in principle) to measure a causal quantity? Does it make sense to define joint probability distributions on quantities that can never be observed together, even in principle? Does it make sense to ask whether race and gender can be 'causes' of an outcome, given that they are not obviously manipulable?

We will discuss some of these from a personal perspective, as well as giving examples of somewhat less careful causal interpretations found in the media.

9.2 Invited - Challenges and insights into Small Island Statistics
Thursday 7 September – 11.40am-1pm

St Helena – Statistics from One of the Worlds Most Isolated Islands

Paula McLeod
Department for International Development

St Helena, a British Overseas Territory in the South Atlantic, is one of the most isolated islands in the world. The speaker, Dr Paula McLeod, was St Helena's Statistician from 2012 to early 2017, responsible for all aspects of the island's statistics.

The island has recently featured UK media due to interest surrounding the construction of the islands spring board to the future- an airport. When Paula and her family arrived on island it required a five day journey from Cape Town by mail ship. At this time the island was better known to many as the place of Napoleon's last exile, Charles Darwins' fertile and attractive emeraldine jewel of the South Atlantic or to mariners as a refuelling station on the pre-Suez canal shipping routes between Europe and Asia.

The territory of St Helena, Ascension and Tristan da Cunha consists of the islands St Helena (on which we will focus), Ascension and the archipelago of Tristan da Cunha. St Helena measures less than 50 square miles (< 100 square km) – about 10 by 5 miles. It is extremely remote, approximately 2000 miles from the South African port of Cape Town and 2200 miles from Salvador in Brazil. The island is home to approximately 4,500 people. The islanders, St Helenians (known as Saints) are a resilient and welcoming ethnically diverse population, descended from British and Dutch sailors and soldiers, liberated African slaves and indentured labourers from China and India.

This session will cover some of the challenges and considerations when producing official statistics for a small island, so far removed from the support from which statisticians typically operate and without the resources and capacity which can easily be taken for granted. It is impossible to talk about working on an isolated island without mention of the joys and challenges of living in one of the most extraordinary places on earth.

9.3 Invited - Water-Energy-Food
Thursday 7 September – 11.40am-1pm

Assessing the utility of composite indicators in analysing the water-energy-food nexus

Scott McGrane, Marian Scott
University of Glasgow

As the global population moves toward 9 billion by 2050, there is a pressing need to ensure a sufficient and secure supply of our critical water, energy and food (henceforth, WEF) resources. Since the 2011 Bonn conference, there has been increasing recognition that our WEF resources are interconnected: water and energy are critical in the production of food, water is a key component of energy generation, and energy is used extensively in the treatment and pumping of water. Actions which impact one of the WEF resources, often result in significant consequences across the others. For example, droughts result in less water being available for food growth or energy production, whilst an increase in fuel prices result in increasing food prices. The capacity of WEF systems to recover from shocks is a measure of their resilience (a concept that has been appropriated from sustainability studies). To measure WEF system resilience, finding a way of representing the WEF system in its component parts and holistically, is necessary. Composite indicators are widely utilised as a way of translating large amounts of complex data into manageable indices that are easily interpretable by policy- and decision-makers. Within the context of the WEF nexus, composite indices are slowly being explored as a method of representing complexity across the three sectors, primarily from a perspective of access and availability. Here, we present the development of a WEF resilience (WEFRes) index for the United Kingdom that highlights how changes in one sphere may impact across others, and discuss how such indicators may break down silos and serve policy and decision-making around the WEF nexus.

9.3 Invited - Water-Energy-Food
Thursday 7 September – 11.40am-1pm

Water, Energy, Food (WEF): Computational Modelling and Interactive Visualisation

Ruth E. Falconer, Ismail Haltas, Daniel Glimour, Liz Varga
School of Design & Informatics, Abertay University

A key component of nexus research is the development of tools that both facilitate assessment and clearly communicate the state of the Water, Energy and Food (WEF) system, and its associated interdependencies and connections. Featuring policy makers as end users, a recent review of nexus tools suggests they might be broadly classified in terms of three objectives: modelling (including optimisation), sustainability assessment and visualisation. In most cases the end users of such tools are assumed to be the principal stakeholders (e.g. those controlling policy or finance) whereas it is increasingly recognised that wider inclusion and engagement of stakeholders promotes transdisciplinary dialogue that can enhance the design and uptake of appropriate interventions. We firstly describe how computational models and visualisation techniques have been used in other domains to promote wider and inclusive decision making. We then propose a methodology for assessing and communicating the effect of socio-technical innovations on the WEF nexus space. This methodology frames multiple questions and scales, to assess the impact of the innovation on WEF nexus sustainability. Finally we discuss how such an approach can deal with the inherent uncertainty of the WEF system.

9.3 Invited - Water-Energy-Food
Thursday 7 September – 11.40am-1pm

Decision support for UK food poverty using coherent inference in integrating decision support systems

Martine J. Barons, Jim Q. Smith
University of Warwick

Decision making in dynamic environments often needs to accommodate vast streams of data and huge models. A subjective expected utility policy making centre managing complex, dynamic systems has to draw on the expertise of a variety of disparate panels of experts and then integrate this information coherently in order to explore and compare the efficacy of different candidate policies. We have developed a formal statistical methodology to network together the diverse supporting probabilistic models needed to achieve this and sufficient conditions that ensure inference remains coherent before and after accommodating relevant evidence into many different classes of such processes. The methodology is illustrated throughout using a decision support systems designed to support policy makers in addressing the complex challenges of food poverty in the UK.

9.4 Invited - Soccer analytics - what's it all about?
Thursday 7 September – 11.40am-1pm

Soccer analytics - what's it all about?

Omar Chaudhuri, Ted Knutson, Rory Campbell

This talk is presented by three individuals who have pioneered the use of statistical analysis in professional football.

First, they describe their path into the industry, providing useful advice to those who are interested pursuing a career in sports analytics. They have all come through different routes, including through blogging and betting, and provide insight into the challenges around being an amateur on the 'outside' of the game.

Then in a panel session, the three discuss the unique environment that is professional football. Though American sports have increasingly utilised data in all areas of decision making, football has been slow to adopt new ways of thinking. Their discussion centres around the models they use within clubs, but more importantly how these models are used, explained and visualised to people who have no background in statistics - and are often very resistant to change.

Finally, they discuss what's the next frontier for use of data in football - and indeed sport. What are the models that need building, and analysis that needs doing, in order to help football clubs win?

9.5 Invited - Large scale inference and modern applications
Thursday 7 September – 11.40am-1pm

Testing One Hypothesis Multiple Times

Sara Algeri, David van Dyk
Imperial College London

When seeking to identify one individual signal over a wide range of possibilities, the cost of a false-positive has the potential to be enormous. In the context of hypothesis testing, this may lead to both stringent significance requirements, which limit the use of simulation and re-sampling methods, and failure of the most commonly used inferential procedures to correct for the multiplicity of tests being conducted.

Statistically, this is as an example of a hypothesis test where a nuisance parameter is present only under the alternative. In this work, we propose a general method to address this problem by combining the outcomes of several dependent tests via a global test statistic, and specify an upper bound for the resulting global p-value which is shown to be less conservative than Bonferroni, equally generalizable, easy to compute, and sharp under long-range independence.

We also show that, under suitable sufficient conditions, it is possible to identify scenarios where the simple Bonferroni correction can be used to provide inferences under stringent significance requirements that are not overly conservative.

9.5 Invited - Large scale inference and modern applications
Thursday 7 September – 11.40am-1pm

Covariance change point in high dimensional settings

Yi Yu
University of Bristol

In this paper, we tackle the high dimensional covariance change point detection problem without extra assumptions on the covariance structure, where $p < n$ and p is allowed to diverge as $n \to \infty$; to be specific, the observations $X_i \in \mathbb{R}^p$, $i = 1, \ldots, n$ are independent sub-Gaussian random vectors with covariance Σ_i, and $\bigl\{\Sigma_i\bigr\}_{i=1}^n$ are piecewise constant. Methods based on binary segmentation and wild binary segmentation are introduced, with the consistency results coincide with those of the mean change point detection problems. In addition, we also propose a variant of wild binary segmentation using random projection, namely wild binary segmentation with random projection, the change point estimator location rate of which is improved and is proved to be optimal in the minimax sense.

9.5 Invited - Large scale inference and modern applications
Thursday 7 September – 11.40am-1pm

Methods of network comparison

Sofia Olhede, Patrick Wolfe, Pierre-Andre Maugis
UCL

The topology of any complex system is key to understanding its structure and function. Fundamentally, algebraic topology guarantees that any system represented by a network can be understood through its closed paths. The length of each path provides a notion of scale, which is vitally important in characterizing dominant modes of system behavior. Here, by combining topology with scale, we prove the existence of universal features which reveal the dominant scales of any network. We use these features to compare several canonical network types in the context of a social media discussion which evolves through the sharing of rumors, leaks and other news. Crucially, our results allow networks to be quantified and compared in a purely model-free way that is theoretically sound, fully automated, and inherently scalable.

9.6 Invited - STEM showcase: Showcasing Statistics Outreach & Public Engagement Activities
Thursday 7 September – 11.40am-1pm

STEM: A Strathclyde Perspective

Johnathan Love, Martin Paton, Kate Pyper, Louise Kelly
University of Strathclyde

In recent times, the importance of engaging children and young adults in Science, Technology, Engineering and Mathematics (STEM) subjects has become a prominent feature of Government, Further Education and Higher Education institution strategies. STEM plays a key role in society and is a major component of future economic growth. There is often, however, a lack of engagement beyond compulsory school level and this has led to a skills gap. In this talk, we introduce a STEM engagement activity that has recently been introduced in the Faculty of Science, University of Strathclyde. The activity is the *Strath Science Scouts* initiative and is based around a key resource within the Faculty – our students.

The Strath Science Scouts initiative involves volunteer undergraduate and postgraduate students studying in the Faculty of Science. They host and organise STEM engagement activities with various schools and year groups. Such activities include working with primary school children in the subjects of Mathematics, Chemistry and Forensic Science via a “Science Mystery” event and exploring aspects of Technology via our “Design an App Challenge”. We also host primary school classes at the University of Strathclyde’s *Fab Lab*, allowing them to interact with various aspects of technology and design. For secondary school pupils, we undertake advanced versions of the events described above, in addition to participating in *Speed Networking*, *Science Cafés* and *Career Stall* based events to promote STEM subjects and the potential career paths in these disciplines. Subject specific events such as interactive Statistics laboratories using Minitab, which allow students studying a school level course in Statistics to apply their skills to real-life data, are also run. Most recently, we have launched a *Digital Mentoring* scheme.

In this talk, these activities will be outlined in detail and ideas for future expansion will be presented.

9.6 Invited - STEM showcase: Showcasing Statistics Outreach & Public Engagement Activities

Thursday 7 September – 11.40am-1pm

Getting hands-on with statistics: ideas without numbers

Simon White
MRC Biostatistics Unit

Statistics is often perceived as a separate thing that needs to be learnt, an impression that can persist from the subject's introduction at school through to higher education. This can cause problems of translating concepts across different contexts to solve problems.

One possible barrier to teaching or demonstrating statistics in an engaging context is a problem of being stuck between requiring mundane/numbing calculations or requiring computers/software. The former is easy to set an activity, the latter is usually beyond the scope of an activity; both of these extremes can obscure the statistical ideas we are trying to teach or communicate.

The Royal Statistical Society's Education Committee remit includes statistical education in practice, promoting careers in statistics, and developing statistical literacy. In line with these aims the Committee is co-ordinating the development of hands-on statistical activities for its Fellows, and STEM Ambassadors within STEMNET, who are invited into careers fairs, schools and science clubs to run sessions or workshops.

This talk will present an approach to designing activities presenting statistical ideas using an interactive (with a physical/tactile component) format to enhance engagement. We will present two example activities, 'how many ducks?' and 'targeting radiotherapy'.

9.6 Invited - STEM showcase: Showcasing Statistics Outreach & Public Engagement Activities
Thursday 7 September – 11.40am-1pm

Penguins, Playing Cards, and Statistics Engagement

Laura Bonnett
University of Liverpool

Mathematics and its applications are vital as numeracy is a core skill for all adults in life generally, a mathematically well-educated population will contribute to the country's economic prosperity, and mathematics is important for its own sake. Unfortunately, maths often receives a bad press, especially from school pupils who may say "maths is boring", or "it's too hard"!

As mathematicians and statisticians, we have a duty to inspire the next generation and to engage the public with mathematics and statistics; to encourage enjoyment of the subject, to enhance and enrich study beyond the curriculum, and to encourage unusual ways of communicating our science. We hope to spread the good word to students and beyond, eventually assisting society to become more mathematically literate so that people's understanding of data, risk, and probability can inform their daily decision making, leading to better outcomes.

The Royal Statistical Society's Education Committee remit includes statistical education in practice, promoting careers in statistics, and developing statistical literacy. In line with these aims the Committee is co-ordinating the development of hands-on statistical activities for its Fellows, and STEM Ambassadors within STEMNET, who are invited into careers fairs, schools and science clubs to run sessions or workshops.

Following on from the successful launch of four hands-on activities at last year's RSS conference, this talk will discuss two new activities which are currently being developed – how many penguins, and sociable cards. "How many penguins?" utilises an aerial photograph of penguin guano ('poo') to investigate the concepts of populations and sampling. "Sociable cards" enables a demonstration of 'statistical magic' courtesy of a variant on Kruskal's Count. The practical and statistical methodology of both hands-on activities will be described enabling attendees to recreate and run these activities for themselves at any relevant future events.

9.7 Invited - Tools that Statisticians need to work as a Data Scientist
Thursday 7 September – 11.40am-1pm

What does a statistician need to know about machine learning?

Aimee Gott
Mango Solutions

The rise of data science has changed the way in which we, as statisticians, need to think and work. New techniques are augmenting those of traditional statistics but they come with challenges for the statistician. From a new vocabulary to an alternative way of model development, there is an increasing need for us to move away from being the traditional focused statistician to the multi-faceted data scientist with a broader look on analytic approaches. But what do we really need to know to survive in this new world of analytics? In this talk we will delve into some of these ideas, highlighting some of the key machine learning algorithms and approaches to modelling so we can adapt to this new world.

9.7 Invited - Tools that Statisticians need to work as a Data Scientist
Thursday 7 September – 11.40am-1pm

Data Science: Engaging with the Business

Richard Pugh
Mango

The "Data Science" movement has led to a rise in popularity of proactive analytics to inform better decision making. However, this shift from analytics as a "reactive" study to a more strategic practice has led to the demand for modellers with a broader skillset. Companies who are striving to be "data driven" are desperate to hire analytic "unicorns" who can code like a software developer and analyse data using an increasingly range of analytic approaches. Perhaps beyond the "technical" skills demanded of a Data Scientist, there is a need for analytic teams to be able to interact with business functions in a more collaborative manner.

This presentation will look at the demands on the modern data scientist from a "business" perspective, including the use of language, the exploration of analytic opportunities, the discuss of data science "success" and storytelling with data.

9.7 Invited - Tools that Statisticians need to work as a Data Scientist
Thursday 7 September – 11.40am-1pm

All the other things a data scientist needs to know

Matthew Upson
Government Digital Service

Data science is a mixed discipline drawing from a number of fields. Whilst a good understanding of statistics and machine learning is critical for the role, equally, data scientists tend to be adept at a range of other skills, many coming from the software engineering world. This presentation explores ‘the other things’ data scientists need to know that are not generally covered in a typical statistical education: from working in an agile team to continuous integration, and (almost) everything in between.

9.8 Invited - Financial Risk modelling
Thursday 7 September – 11.40am-1pm

Accounting for heterogeneity and macroeconomic variables in the estimation of transition intensities for credit cards

Vinai Djeundje, Jonathan Crook
Credit Research Centre, University of Edinburgh

The literature has considered intensity models that give predictions of the probability, for each customer, that he/she will transit from one state of delinquency to another between any two months in the life of the loan. The transitions include not only transitions into further delinquency but also transitions to lesser states of delinquency, that is cure. We now extend this work by treating time as discrete, including a frailty term relating to the individual cases, including macroeconomic variables and represent the baseline intensity as B-splines. The inclusion of frailty means that any statistical bias that may exist because of the omission of unobserved individual level effects should be removed. We show the likelihood function and apply the method to a large dataset relating to credit card accounts. We conclude that the use of B-splines allows the detection of noticeably different baseline hazards between the jump processes, that the inclusion of the random effects is supported by highly significant variances of these effects for all models, that the inclusion of frailty generally reduces the significance of the covariates and when flexible baseline specifications are used the inclusion of random effects does not increase predictive accuracy.

9.8 Invited - Financial Risk modelling
Thursday 7 September – 11.40am-1pm

Default contagious risk assessment though space-time point processes

Giada Adelfio, Marcello Chiodi, Paolo Giudici
Università di Palermo

The global financial crisis and, more recently, the European sovereign crisis, have led to an increasing research literature on systemic risk, with different definitions and measurement models.

An important distinction between models derives from the use of a cross-sectional rather than a time-dynamic perspective: while the former mostly concentrates on contagion between the institutions operating in the market, the latter focuses on the generating cause-and-effect relationships over time.

While contagion models identify transmission channels, thus describing how a crisis spreads through the whole system, time-dependent models associate a specific risk measure to individual institutions, with the aim of predicting what will happen to them in the nearby future, in an early-warning perspective.

In this paper we propose a model that can accommodate for both perspectives, by means of space-time point processes.

The application of the methodology to the CDS spread time series of 54 European financial corporates shows that the proposed model provides important insights in terms, not only from a descriptive, but also from a predictive viewpoint.

9.8 Invited - Financial Risk modelling
Thursday 7 September – 11.40am-1pm

Spatial contagion in mortgage defaults: a Bayesian survival approach

Raffaella Calabrese, Jonathan Crook
Business School University of Edinburgh

This paper proposes a spatial discrete survival model to estimate the time to default for UK mortgages. The model includes a flexible parametric link function given by the Generalised Extreme Value Distribution and a dynamic spatially varying baseline hazard function to capture neighbourhood effects over time. We incorporate time varying variables into the model and we estimate it in a Bayesian framework using Gibbs sampling. The gains of the proposed model are illustrated through the analysis of a UK dataset.

10.1 Contributed – Medical Statistics: Biostatistics
Thursday 7 September – 2pm-3pm

Longitudinal analysis strategies for high-dimensional phenotypes

James Staley, Andrew Simpkin, Matthew Suderman, Kate Tilling
University of Bristol

Background

High-dimensional phenotypes such as epigenetics, metabolomics and proteomics vary over time, and modelling their trajectories could improve our understanding of disease mechanisms. However, the computational cost of fitting multi-level models to high-dimensional data has limited longitudinal analyses to focus on certain markers (e.g. a subset of 450,000 CpG sites) found to be significantly associated through cross-sectional testing (e.g. CpG sites associated at baseline) as opposed to fitting longitudinal analyses for all markers.

Methods

We propose using linear regression across the repeated measures, estimating clustered robust standard errors via an extended version of the standard sandwich estimator, as a less computationally intensive strategy than multi-level modelling. This approach has the added benefit of more consistent convergence than multi-level models. Using simulations, we compared these two longitudinal approaches, as well as two approaches based on cross-sectional testing (significantly associated at the first time-point only and at any time-point), for identifying epigenetic change over time related to an exposure. We then applied all approaches to repeatedly measured blood DNA methylation profiles from the Accessible Resource for Integrated Epigenomics Studies (ARIES) in relation to prenatal smoking.

Results

Simulations: Restricting association testing to the first time-point alone identified fewer associations than performing association analyses at each time-point or applying the longitudinal modelling approaches to the full dataset. In spite of being >30 times faster than multi-level models, linear regression models with clustered robust standard errors identified similar sets of associations (>95% agreement) with almost identical effect estimates.

Application: Both longitudinal modelling approaches identified comparable sets of methylation sites in ARIES related to prenatal exposure to smoking, including additional sites not associated in analyses only at baseline or at individual time points.

Conclusion

Linear regression with clustered robust standard errors is an appropriate and efficient approach for longitudinal analysis of high-dimensional data.

10.1 Contributed – Medical Statistics: Biostatistics
Thursday 7 September – 2pm-3pm

Probabilistic deconvolution of microarray data using a hierarchical Bayesian model

Daniel Kennedy, Nicole White, Kerrie Mengersen, Rodney Lea, Miles Benton, Lyn Griffiths
Queensland University of Technology (QUT)

Microarray data are often derived from samples of heterogeneous tissue, which are comprised of several different cell-types. This is represented statistically as a convolution, where cell-types can be thought of as source variables and the array data are a mixture of these sources. When a disease state or phenotype may only affect a single cell-type, there is interest in performing inference on specific sources within the convolution. To address this problem, we develop a hierarchical Bayesian model to perform source-specific inference in a convolution. The approach leverages prior information from cell-sorted data, and regularizes over the source effects and over neighboring loci. The model also offers the flexibility to aggregate different types of microarray data into the same analysis and to make a combined inference in locating phenotypically important sites and estimating effects. The model was applied to a simulated data set for validation, then to samples of DNA methylation derived from a heterogeneous mixture of blood cell types, in order to make phenotype inferences for specific cell-types.

10.1 Contributed – Medical Statistics: Biostatistics
Thursday 7 September – 2pm-3pm

Bayesian modelling of disease evolution in musculoskeletal epidemiology

Amelia Green, Gavin Shaddick, Alison Nightingale, Rachel Charlton, Julia Snowball, Catherine Smith, Neil McHugh
University of Bath

Psoriatic arthritis (PsA) is a progressive and often destructive joint disease affecting approximately 20% of patients with psoriasis. The inflammation, joint damage and deformity associated with PsA adds substantially to the disease burden for psoriasis patients, leading to an impaired quality of life. Understanding the role of risk factors, such as obesity, is essential in determining which patients are at the greatest risk of developing PsA. The ability to utilise information from routinely collected data contained in databases such as the Clinical Practice Research Datalink (CPRD), which contains medical records for ca. 11.7 million individuals, has the potential to greatly increase this understanding. However, there may be issues when using databases for purposes beyond that for which they were originally designed. In a traditional epidemiology study, data would be collected in accordance with a specified study design, for example recording measurements of risk factors, such as body mass index (BMI), at specified points in time. This is not the case in the CPRD where, from a statistical point of view, the non-routine collection of data results in missing values. Bayesian hierarchical models (BHM) provide a coherent framework within which the inherent uncertainty present in clinical measurements, and the estimation of missing data, can be propagated throughout the modelling process and incorporated in resulting estimates of risk and associated measures of uncertainty, such as confidence intervals. A case study using data extracted from the CPRD is presented in which BHMs were used to estimate the risk of developing PsA in patient with psoriasis. Using data on 90,182 psoriasis patients, the possibility of lagged and cumulative risks associated with changing BMI were investigated. Results showed that increased BMI was significantly associated with a higher risk of PsA, with evidence that reductions in BMI over time reduce risk.

10.2 Contributed – Official statistics and public policy: Methodology
Thursday 7 September – 2pm-3pm

Seasonal Adjustment Components for Time Series: An Alternative Estimate for Excess Winter Mortality

Atanaska Nikolova, Duncan Elliott, Paul Smith
Office for National Statistics

Mortality levels are typically higher in winter, and Excess Winter Mortality (EWM) is an annual indication of this additional number of winter deaths used by the Office for National Statistics (ONS). Obtaining an accurate measure of this construct is important in order to understand links with social and environmental factors, and also for the purpose of designing policy and interventions to decrease preventable deaths. The current method of estimating EWM compares the number of deaths during a four-month winter period (December-March) with the average of deaths for the remaining non-winter months. Therefore, a potential limitation is that atypical mortality patterns in non-winter months can lead to over-estimation or under-estimation of EWM. The current project aims to explore an alternative method of estimating EWM. Rather than using mortality data in its raw format and adopting non-winter months as a comparison baseline, the data are decomposed into three components which are estimated as part of the seasonal adjustment process for time series. These components reflect a trend-cycle, seasonal variation, and a residual irregular element. The latter captures unsystematic fluctuations in the underlying pattern. As such, the irregular element can be used to derive a purer indication of EWM. It is also a useful indicator for cross-county comparisons. This approach also has the benefit of not needing a full twelve-month span of data in order to be estimated, since this can be done on a rolling basis as more data becomes available.

10.2 Contributed – Official statistics and public policy: Methodology
Thursday 7 September – 2pm-3pm

Back to the Source: the realities of data collection

Daria Gromyko
HMRC

Official Statistics come in all shapes and sizes and from a wide variety of sources, among which surveys and administrative data are the big players. Each comes with its own pros and cons, but in each case we seek to draw accurate conclusions from data often afflicted with quality issues.

In recent years much emphasis has fallen on administrative data as being more readily available and cheaper than survey data. More computing power and the dawn of Data Science have driven a hype around the exploitation of Big Data - but is all administrative data in its current state suited to this approach? And are there some things that good old-fashioned questionnaire and survey design principles and methodology can contribute to the Big Data age?

This talk will focus on Compliance systems in HMRC as the primary example of collecting administrative data, touching on other examples from the Home Office and the Department for Education to illustrate some practical aspects of data collection processes, and how we as statisticians can influence these in large organisations such as Government departments, in order to achieve better quality data and more impactful analysis.

10.2 Contributed – Official statistics and public policy: Methodology
Thursday 7 September – 2pm-3pm

Ecological inference with distribution regression: kernel methods to model voting patterns in US presidential elections with individual-level demographic data

Seth Flaxman
Oxford

We combine fine-grained spatially referenced census data with the vote outcomes from the 2012 and 2016 US presidential election. Using this dataset, we perform ecological inference using distribution regression (Flaxman et al, KDD 2015) with a multinomial-logit regression so as to model the vote outcome Trump, Clinton, Other / Didn't vote as a function of demographic and socioeconomic features. Ecological inference allows us to estimate "exit poll" style results like what was Trump's support among white women, but for entirely novel categories. We also perform exploratory data analysis to understand which census variables are predictive of voting for Trump, voting for Clinton, or not voting for either. All of our methods are implemented in python and R and are available online for replication.

Joint work with Dougal Sutherland, Yu-Xiang Wang, Yee Whye Teh, and Alex Smola. Pre-print: https://arxiv.org/abs/1611.03787

10.3 Contributed - Environmental/Spatial statistics: Spatial Statistics
Thursday 7 September – 2pm-3pm

Detecting Change: Development of a new tool for spatially explicit power analysis

Lindesay Scott-Hayward, Monique Mackenzie
University of St Andrews

New construction developments may impact animals which use the development site; for example, with an offshore windfarm marine mammals or seabirds may avoid the site entirely or redistribute themselves within the site. The challenge is to determine if any changes in abundance and distribution are due to a development (either directly or indirectly) or if these changes would have occurred anyway in the absence of any development. Surveys of the site are thus generally conducted before any development takes place, during construction and after construction in order to reliably determine any effects. Statistical methods can be used to identify both temporal and spatial changes at the site, but the ability of a study to detect genuine change, power, quantifies the chance that a study will correctly identify genuine change. Historically, power analyses have often been simple and only establish the power to detect site-wide changes in overall numbers rather than focus on spatially explicit changes in distributions.

To realistically establish spatially explicit power, a simulation based approach was used and this approach relies on realistically generated data. Specifically, the generated data must represent the signal underlying the process of interest, the variability observed in the data and patterns that could not be explained by covariates used in the data generation model (e.g. which present as residual correlation).

Subsequent to the generation of data simulated for the power analysis can be undertaken. The method involves inducing changes to the generated baseline data; fitting model(s) to the new sets of post-change generated data and ultimately quantifying the power to detect change (across the survey area) along with sense-checking the results. We present the methods for this approach and some spatially explicit power results using pre-construction data from a large offshore wind farm (Lincs) as a case-study.

.

10.3 Contributed - Environmental/Spatial statistics: Spatial Statistics
Thursday 7 September – 2pm-3pm

Age-Period-Cohort Analysis Using Spatially-varying Coefficients With Application to U.S. Mortality, 1999-2015.

Pavel Chernyavskiy, Mark P. Little, Philip S. Rosenberg
National Cancer Institute/National Institutes of Health

Age-period-cohort models are widely used to decompose trends in population rates into age, calendar period, and birth cohort effects. Despite their popularity, methodology to fit parsimonious age-period-cohort models to data in geographically-organized regions (e.g., U.S. states) is not well developed. Here, we present a general Bayesian Markov Chain Monte Carlo method for modelling trends in population rates collected across adjacent geographic regions. We employ a spatially-varying intercept, age trend, and birth cohort trend to capture between-region heterogeneity and spatial correlation. In addition to spatial correlation, our method allows the random intercept and random trends to be correlated, as is often the case in mixed effects models. We do this by adapting the Generalized Multivariate Conditionally Auto-regressive model to specify the joint distribution of the three random effects. Estimation is carried out using a Gibbs sampler in R. We apply our model to U.S. state-level mortality in young (aged 25-50) white non-Hispanic men and women to assess the impact of fatal drug overdoses (OD) on total mortality over the period 1999-2015. In this population, OD accounted for 15.6% of deaths in women, and 14.0% of deaths in men. We show that OD was responsible for much of the rising mortality among both men and women, and that age-adjusted trends are clustered in parts of the Deep South and Appalachian regions of the U.S. We further demonstrate that the largest increases tend to occur where baseline risk is also elevated, indicating that disparities among states have grown over time. In summary, we have developed a general and easily fitted model that can be used to assess differences in trends among potentially-contiguous geographic regions. Our model can be readily applied to other outcomes (e.g., cancer incidence), and at a different geographic scale (e.g., county).

10.3 Contributed - Environmental/Spatial statistics: Spatial Statistics
Thursday 7 September – 2pm-3pm

On some local estimation problems in large scale spatial data analysis

Sucharita Ghosh, Andri Baltensweiler, Markus Huber, Gabrielle Moser, Carlos Ochoa
Statistics Lab, Swiss Federal Research Institute WSL, Birmensdorf, Switzerland

We consider analysis of spatial observations from several thousand locations in a large area. Examples include forest inventory data on a nationwide grid of permanent plots, satellite data on a global scale, household data from large cities, and so on. When very large spatial scales are involved, distributional parameters may be functions of geographical coordinates and not remain as constants across the entire spatial scale. This is a well-known problem and one approach is to consider local estimation, e.g. nonparametric or semiparametric estimation via kernel smoothing. We illustrate this using two selected problems: (a) Fitting zero-inflated distributions to counts, with or without auxiliary variables. This problem occurs when there are too many zero observations which cannot be explained by parametric models like the Poisson. (b) Detecting spatial locations where a smooth nonparametric surface exceeds a threshold, as for instance the local mean value being too high or too low. We address relevant technical issues and work through some numerical examples.

10.4 Contributed - Applications of Statistics
Thursday 7 September – 2pm-3pm

Effects of national housing quality standards on hospital emergency admissions: a quasi-experiment using data linkage

Damon Berridge, Sarah Rodgers, Wouter Poortinga, Rowena Bailey, Rhodri Johnson, Robert Smith, Ronan Lyons
Swansea University Medical School

A housing improvement programme was delivered through a local authority to bring nearly 9000 homes up to the Welsh Housing Quality Standard (WHQS). Homes received multiple elements, including new windows and doors, and heating and electrical systems, through an eight-year rolling work programme. The study aimed to determine the impacts of the different housing improvements on hospital emergency admissions for all residents.

Intervention homes, council homes that received at least one element of work, were data linked to individual health records of residents. Counts of admissions relating to respiratory and cardiovascular conditions, and falls and burns, were obtained retrospectively for each individual in a dynamic housing cohort (January 2005 - March 2015).

The intervention cohort criterion was for someone to have lived in any one of the intervention homes for at least three months within the intervention period. Counts were captured for up to 123 consecutive months for 32,009 individuals in the intervention cohort and analysed using a multilevel approach to account for repeated observations for individuals, nested within geographic areas.

Negative binomial regression models were constructed to determine the effect for each element of work on emergency admissions for those people living in homes in receipt of the intervention element, compared to those living in homes that did not meet quality standards at that time. We adjusted for background trends in the regional general population, as well as for other confounding factors.

People of all ages had fewer admissions for cardiovascular and respiratory conditions, and fall and burn injuries while living in homes when the electrical systems were upgraded, compared to the reference group. Reduced admissions were also found for new windows and doors, wall insulation and garden paths. In summary, improving housing to national standards reduces the number of emergency admissions to hospital for residents.

10.4 Contributed - Applications of Statistics
Thursday 7 September – 2pm-3pm

The Type Ia Supernova Color-Magnitude Relation and Host Galaxy Dust: A Simple Hierarchical Bayesian Model for Cosmology

Kaisey Mandel
Harvard University

Type Ia supernovae are faraway exploding stars used as ``standardizable candles'' to determine cosmological distances, measure the accelerating expansion of the Universe, and constrain the properties of dark energy. Inferring peak luminosities of supernovae from distance-independent observables, such as the shapes and colors of their light curves (time series), underpins the evidence for cosmic acceleration. Supernovae with broader, slower declining optical light curves are more luminous ("broader-brighter") and those with redder colors are dimmer. But the "redder-dimmer" color-luminosity relation widely used in cosmological supernova analyses confounds its two separate physical origins. An intrinsic correlation arises from the physics of exploding white dwarfs, while interstellar dust in the host galaxy also makes supernovae appear redder and dimmer (extinguished). However, conventional supernova cosmology analyses use a simplistic linear regression of magnitude versus color and light curve shape, which does not model intrinsic supernova variations and host galaxy dust as physically distinct effects, resulting in unusually low color-magnitude slopes. I constructed a probabilistic generative model for the dusty distribution of extinguished absolute magnitudes and apparent colors as the convolution of an intrinsic supernova color-magnitude distribution and a host galaxy dust reddening-extinction distribution. If the intrinsic color-magnitude slope differs from the host-galaxy dust law, this convolution generates a specific curve of mean extinguished absolute magnitude vs. apparent color. I incorporated these effects into a hierarchical Bayesian statistical model for supernova measurements, and analyze an optical light curve dataset comprising 248 nearby supernovae at $z < 0.10$. The conventional linear fit obtains an effective color-magnitude slope of 3. My model finds an intrinsic slope of 2.3 ± 0.3 and a distinct dust law of R_B = 3.8 ± 0.3, consistent with the average for Milky Way dust, while correcting a systematic distance bias of ~0.10 mag in the tails of the apparent color distribution. https://arxiv.org/abs/1609.04470

10.4 Contributed - Applications of Statistics
Thursday 7 September – 2pm-3pm

Joint modelling of financial and economic cycles: Recent research findings with an application to the Euro area

Dan A. Rieser, Monica Billio, Roberto Casarin, Marcella Lucchetta, Antonio Paradiso
European Commission

Since the global financial and economic crisis, researchers and academics started paying increasing attention to the relevance, role characteristics and main features of the financial cycle. As a result, research in this area has started focusing on the analysis of the relationships between financial and economic cycles in statistical terms in order to better describe the cyclical movements of the economy and to obtain a more effective system for anticipating or detecting in real time future crises. Nonetheless, a coherent definition of "financial cycles" and a rigorous assessment of the interlinkage between financial cycles is still missing.

The purpose of this paper is to provide an update on recent work that EUROSTAT has carried out in this area. Based on a assessment of the theoretical aspects of the financial cycle, early results from an empirical analysis of the Euro area financial cycle are provided with particular emphasis on the relationships between financial and economic cycles in the Euro area and selected countries.

To achieve the joint modelling of financial and economic cycles, a Markov Switching processes has been adopted to study the synchronization of the country-specific cycles and the aggregate financial cycle. Early research findings from the indicate that this approach is suitable for modelling regional business cycles and the financial cycle to assess the importance of the financial and the EU business cycle synchronization. Nonetheless, more work remains to be done to promote a better understanding of the precise nature of synchronicity between financial and economic cycles and to measure the degree of concordance of each of the countries under observation.

On choosing mixture components via non-local priors

Jairo Alberton Fuquene Patino, Mark Steel, David Rossell
University of Warwick

Choosing the number of components remains a central but elusive challenge in mixture models. Traditional model selection criteria can fail to enforce parsimony or result in poorly separated components of limited practical use. Non-local priors (NLPs) are a family of distributions that encourage parsimony by enforcing a separation between the models under consideration. We formalize NLPs in the context of mixtures and show how they lead to extra parsimony and well-separated components that have non-negligible weight, hence interpretable as distinct subpopulations. We derive tractable expressions and suggest default prior settings aimed at detecting multi-modal densities. We also give a theoretical characterization of the sparsity induced by NLPs and propose easily implementable algorithms to obtain the integrated likelihood and parameter estimates. Although the framework is generic we fully develop the multivariate Normal mixture case based on a novel family of exchangeable moment priors. The proposal is illustrated using simulated and real data sets. Our results show a serious lack of sensitivity of the Bayesian information criterion (BIC) and insufficient parsimony of local prior and shrinkage counterparts to our formulation, which strikes a reasonable balance between power and parsimony.

10.5 Contributed - Methods and theory: Bayesian Inference
Thursday 7 September – 2pm-3pm

Eliciting and specifying multivariate prior distributions using vines

Kevin Wilson
Newcastle University

In this talk, we consider the problem of the specification of an informative prior distribution, or more generally an uncertainty distribution, for multiple dependent unknown quantities. We consider a specific instance of this general problem, specifying a prior distribution for the probabilities in a multinomial model. We utilise vine copulas: flexible multivariate distributions built using bivariate copulas stacked in a tree structure. We take advantage of a specific vine structure, called a D-vine, to separate the specification of the multivariate prior distribution into that of marginal distributions for the probabilities and copula parameters for the bivariate copulas in the vine. We propose a method to find the most suitable vine structure given the expert's judgements based on maximising the dependency in the first tree of the D-vine. Suitable methods to elicit all of the parameters in the vine are proposed, rooted in asking questions based on observable quantities. We embed the elicitation within the Sheffield Elicitation Framework. The approach is illustrated using an example from a desensitised industrial case study involving the condition of an ageing engineering structure. We demonstrate the ability of experts to specify the required quantities in an elicitation exercise with a school teacher in England.

10.5 Contributed - Methods and theory: Bayesian Inference
Thursday 7 September – 2pm-3pm

Inference in complex systems using multi-phase MCMC sampling with gradient matching burn-in

Alan Lazarus, Dirk Husmeier, Theodore Papamarkou
University of Glasgow

Statistical inference in nonlinear differential equations (DEs) is challenging. The log likelihood landscape is typically multimodal, and every parameter adaptation requires a computationally expensive numerical integration of the DEs. When one adopts a Bayesian approach to the problem, the required numerical approximations ensure the inefficiency of the adopted inference scheme. This computational complexity can be overcome by obtaining a representative objective function that manages to capture the fundamental property of the likelihood landscape—the global optimum. By first obtaining a smooth interpolant to the data, we can implement a gradient matching objective function that, instead of quantifying how well the solution of the DEs matches the data, quantifies how well derivatives predicted by the DEs match derivatives obtained from the interpolant to the data. Obviously, since we are no longer sampling from the true likelihood but instead a cheap proxy, bias is introduced to the inference problem. Current research focuses on reducing this bias by introducing a regularising feedback mechanism from the DEs back to the interpolation scheme. The idea is to make the interpolant maximally consistent with the DEs. Although this paradigm has proven to improve performance over naive gradient matching, the feedback loop fails to fully negate bias in the final estimate. A natural progression would be to implement a computationally cheap surrogate in the discarded burnin steps before making use of the true likelihood function in the sampling phase of the algorithm. Assuming this hypothesis, we propose the use of a three-phase technique in which a cheap surrogate likelihood is used in the initial burnin phase alone. Through examples possessing multimodal likelihoods, we will show the ability of the algorithm to avoid local entrapment whilst efficiently obtaining accurate parameter estimates.

Acknowledgements: This project was supported by a grant from the Engineering and Physical Sciences Research Council (EPSRC) of the UK, grant reference number EP/L020319/1.

10.6 Contributed - Communicating and teaching statistics: Communication of statistics with data visualisation
Thursday 7 September – 2pm-3pm

Teaching basic statistics for social worker students, using Learning-by-doing pedagogy.

Ingrid Svensson
Department of Statistics, Umeå University

The program for social workers at Umeå University, Sweden, includes a mandatory course in statistics. Many students lack mathematical skills and confidence when it comes to learning statistics, and we have therefore developed the course in a learning-by-doing framework. The students have access to short films covering the different parts of the course, follow-up lectures where we use an audience response system, and lessons where the students discuss and solve problems together. A central part of the course is a group assignment in which the students perform a survey study, using the class itself as the population of interest. The idea with using a practical assignment is to give the students a real-life experience, which we think is a favorable way to support them in their learning process. The assignment is backed up by several seminars, which involve presentations in small inter-groups where the students give feedback to each other. The first seminar treats the purposes and aims of their studies, and the groups present their plans for the study in inter-groups in a poster session. The second seminar is a very practical seminar where we work with the students' questionnaires by first answering them individually, and then analyzing the questions together in smaller groups before the students distribute them to a random sample of the class. In a third seminar, the students analyze their collected data, and discuss their descriptions and visualizations of their data. After this seminar, each student is to prepare an own short video presentation (often power point based), which also is discussed and given feedback upon in a fourth seminar. The set up of the course has successfully been used in on-line courses as well as in Campus courses, and in the course evaluations, the students say that they appreciate the framework of the course.

10.6 Contributed - Communicating and teaching statistics: Communication of statistics with data visualisation
Thursday 7 September – 2pm-3pm

Markov Processes: Dynamic visualizations to enhance student understanding

Stephanie Budgett
University of Auckland

Finding ways to improve introductory undergraduate students' understanding of probability ideas and theory is a goal of many first-year university probability courses. In this presentation I explore the potential of a prototype software tool for Markov processes. The tool uses dynamic visualizations to develop in students a deeper understanding of the equilibrium and hitting times distributions.

The tool and accompanying tasks were piloted on six introductory probability students using a two-person protocol whereby students worked in pairs and could discuss with one another proposed actions and what they were thinking.

The talk will focus on these students' interactions with the Markov processes tool. The main findings of this exploratory study suggested that the tool and tasks have the potential to enhance students' probabilistic reasoning. The tool seemed to assist students to:

- engage with and develop some intuition for Markov processes,
- to enhance their distributional ideas,
- to work flexibly between representations, and
- to see mathematical structure.

10.6 Contributed - Communicating and teaching statistics: Communication of statistics with data visualisation
Thursday 7 September – 2pm-3pm

Online automated assessment using R

Deirdre Toher
University of the West of England

Online testing has moved beyond the realm of multiple choice. This talk will introduce a system that allows us to create individualised assessment for students. Using R to generate data and produce analysis, these can then be linked to create unique datasets and results for each attempt. This enables students to attempt the assessment multiple times and get a unique dataset, and set of answers, for each attempt.

This means that this system is particularly suited to use for both formative and summative assessment.   By linking the analysis directly to R output, we can specify numeric answers as well as the traditional multiple choice style of questions traditionally associated with automated online testing.

Examples using simple linear regression, multiple linear regression and generalised linear modelling will be shown.

10.7 Contributed - Methods and theory: Linear Models
Thursday 7 September – 2pm-3pm

Focussed model averaging in generalised linear models

Chuen Yen Hong, David Fletcher, Matthew Parry
Department of Mathematics and Statistics, University of Otago, NZ

Under the frequentist framework, a model-averaged point estimator is a weighted mean of the estimates from each of the candidate models. Focussed model averaging is a method which tailors the weights to a specific parameter of interest. Traditional model averaging procedure such as the AIC, weighs the model based on a measure of its predictive ability regardless of the focus parameter of interest. Considering generalised linear models, we propose a focussed model averaging procedure such that model weights are chosen to minimise an estimate of the mean squared error of the focus parameter. In contrast to existing literature, our proposed method does not rely on a local misspecification assumption. The results are conceptually simpler and are straightforward to apply for generalised linear models. Simulation studies show that our proposed method can be a simpler alternative to existing focussed model averaging procedure and in some cases, outperforms the existing method.

10.7 Contributed - Methods and theory: Linear Models
Thursday 7 September – 2pm-3pm

Identifying variables underlying multicollinearity

Zillur Shabuz, Paul Garthwaite
Open University, UK

Multicollinearity is a common problem in multiple regression analysis. The presence of multicollinearity will inflate the variances of parameter estimates and can radically affect the values of the estimates. We discuss several methods for the diagnosis of multicollinearity and also discuss one new procedure and two older procedures for identifying the collinear sets. The new procedure is based on transformations that partition variance inflation factors into contributions from individual variables, thus providing detailed information about the collinear sets. The procedures are compared using three examples from published studies that addressed issues of multicollinearity using the older procedures.

10.7 Contributed - Methods and theory: Linear Models
Thursday 7 September – 2pm-3pm

Optimal Design, Lagrangian and Linear Model Theories: a Fusion.

Ben Torsney
University of Glasgow

We consider the problem of optimizing a criterion φ(p) subject to several equality constraints: Ap=b. (wlog b ≥0.)

Lagrangian Theory requires that at an optimum all partial derivatives be exactly linear in a set of Lagrange Multipliers.

Thus the partial derivatives, viewed as response variables, must exactly satisfy a Linear Model with the Lagrange Multipliers as parameters. This then is a model ‘without’ errors, implying a ‘fitted model’ with zero residuals.The residuals appear to play the role of directional derivatives, as defined in the optimal design arena when A = $\underline{1}^T$, b=1.

Further we extend a class of multiplicative algorithms, designed to find the optimum in the latter case, to our general problem. This algorithm has two steps:

(i) a multiplicative one, multiplying the current values of the components of p by an positive increasing function of partial or directional derivatives;

(ii) a scaling step under which the products formed in (i) are scaled to meet the summation to one equality constraint:

$\underline{1}^T$p=1.

Step (i) readily applies to our more general problem, while the more challenging

step (ii) has been surmounted.

Results arising from examples of data forming multidimensional contingency tables with constraints on multinomial models defined by fixed marginal distributions or hypotheses of marginal homogeneity will be reported.

10.8 Contributed - Industry and commerce: Trains, Job quality and Data Scientists
Thursday 7 September – 2pm-3pm

The role of a Data Scientist: An Industry Perspective

Wayne Jones
Shell

What does it mean to be a Data Scientist in Industry? Current practitioners of Data Science do not have a pure “Data Science” educational background, principally, because the discipline didn’t exist when such individuals were learning their respective trades in University. As such, many traditionally trained statisticians have transitioned to fill the demand in this new discipline.

In this talk, one such statistician presents a series of industrial data science use cases:

1. Forecasting wind power generation.
2. Optimising inventory stocking levels.
3. Aviation fuel demand forecast modelling.
4. Market basket analysis applied to the Lubricants sales business.

The critical key strengths that statisticians already possess are discussed whilst highlighting potential areas of development to become successful in this new and exciting field.

10.8 Contributed - Industry and commerce: Trains, Job quality and Data Scientists
Thursday 7 September – 2pm-3pm

Estimating job quality using a 'Page Rank' algorithm

Paul Jones
Leeds Beckett University

Economists measure “job quality” in relation to the distribution of employment. This has spurned a literature linked to policy relevant debates on such phenomena as ‘job polarisation’ and the ‘disappearing middle’ of the job market. The system for ranking jobs has been done typically based on ranking occupations by average earnings, following Goos and Manning (2007). Although reasonable, the choice of metric (earnings) is arbitrary. ‘Job quality’ is a multi-dimensional concept and it is not only higher pay which makes us happy.

This paper proposes an alternative way of establishing job-rank utilising the ‘Page Rank' algorithm. Introduced by Sergey Brin and Larry Page (1996), this established a method for ranking webpages based on traffic from one webpage to another, later adopted by Google. The transition matrix based on web-traffic volume from one webpage to another can in turn reveal a steady state vector associated with (ergodic) page ranking.

The algorithm can be applied to employment transitions, based on the probability and direction of job moves from one occupation to another. If people move jobs to better themselves (and voluntary job-to-job moves dominate involuntary job-to-job moves) then we might infer that the destination-job is revealed-preferred to the origin-job. Constructing a transition matrix of job moves can then reveal a ranked distribution of occupations, based, notionally, on where people would like to be in a frictionless and timeless job market.

The paper applies this thinking to construct job quality rankings using employment transition data from the UK Labour Force Survey (LFS) longitudinal data, based on data based on Standard Occupational Classification (SOC 2010), at various levels of disaggregation of occupation. The derived measures are compared to standard measures of job quality based on earnings, and other occupational characteristics. The results provide useful insights and a potential directions for thinking about job quality.

10.8 Contributed - Industry and commerce: Trains, Job quality and Data Scientists
Thursday 7 September – 2pm-3pm

Solving the Location Problem to Enable Rolling Stock Fault Analysis and Key Performance Indicator Creation Using Statistical Modelling

Nick Wray, Stephanie Coates
Peak Statistics

Modern railway vehicles carry an OTMR (On Train Monitoring Recorder) device, which records many channels of data giving information about the current state of a locomotive, such as its speed, acceleration, throttle position, brake setting, cylinder pressure and so on, at very small consecutive time intervals (typically down to 0.1 seconds).

This data can be used to create Key Performance Indicators (KPIs) to monitor the performance of individual trains and direct the most efficient use of maintenance resources.

However, surprisingly perhaps, OMTR data does not include information about the train's location, and, in order to remove the lurking variable "track location" from the analysis, as performance comparisons have to be over the same stretch of track, the train's position has to be inferred from other OTMR data.

The objective is to get rid of the lurking variable "track location", thus allowing analysis of train performance data such as speed, acceleration and so on, and also then allowing the elimination of further lurking variables such as weather conditions, driver identity and so on.

This paper uses data collected from a regional rail network, which consists of lines radiating out from an urban centre. The trains run eight scheduled routes, which may overlap. There are two topological types of route. One is a simple thither and hither linear shuttle, and the other is a run into the centre, into a one-way loop and then a run back out again. Each route has a characteristic number of station calls.

This paper discusses ways of tackling the "location problem", using data vectors such as the timing pattern of door opening signals, which should give a "fingerprint" pattern of the stations, the mileage and timing differences between stopping points (which may be signals and not stations), and the actual number of stops made.

10.9 Invited - Prize winners: Young Statisticians Meeting 2016
Thursday 7 September – 2pm-3pm

Avoiding pitfalls when combining multiple imputation and propensity scores

Emily Granger, Jamie Sergeant, Mark Lunt
University of Manchester

Propensity score (PS) methods are used to minimise confounding when estimating effects of exposures on outcomes. Over the past decade there has been a substantial increase in the use of PS methods in social sciences and health research. They are most widely used in medical research to analyse observational data. Observational studies are prone to bias due to both confounding and missing data. While PS methods are useful to deal with the first problem, multiple imputation is often used for the latter. Unfortunately, it is not known how best to proceed when both techniques are required. There are two methods of combining propensity scores with multiple imputation in the current literature. This research investigates whether the two methods lead to differences in the accuracy or precision of exposure effect estimates.

Both methods start by imputing each missing value multiple times. Propensity scores in each of the resulting datasets are then estimated. In Method 1 the propensity scores are averaged and the average propensity score is used for subsequent analysis. Alternatively, Method 2 uses the propensity scores individually to obtain multiple estimates of the exposure effect which are combined to produce an overall estimate. These methods were compared by conducting an extensive series of Monte Carlo simulations, where the confounders different in number, strength of association with outcome and exposure and direction of association.

Consistently across simulations, Method 2 produced unbiased results with appropriate confidence intervals. Method 1 not only underestimated the uncertainty about the estimate (as might be expected) but also produced biased estimates.

Researchers are encouraged to implement Method 2 when conducting a propensity score analysis with incomplete data. This result is particularly relevant in health and social sciences, where missing data are unavoidable and propensity scores are on the rise.

10.9 Invited - Prize winners: Young Statisticians Meeting 2016
Thursday 7 September – 2pm-3pm

An introduction to the Longitudinal Study of Young People in England

Viktoria Vianeva
Department for Education

The second Longitudinal Study of Young People in England (LSYPE2) is a large scale research project led by the Department for Education. The main objective of the study is to understand the changing lives of young people in England, to ensure that government policies and services are informed by the experiences of young people in England.

The survey follows a cohort of young people each year from when they are 13/14 up to 19/20 and covers a wide range of topics such as attitudes to education and employment, subject choice, family relationships, mental health and many others. The longitudinal nature of the survey allows for interesting analyses looking at transitions over time and identifying predictors of different outcomes. Data linkages to administrative data sources, such as the National Pupil database, further enhance the value of the survey. Being the second cohort of the same study, we are also able to make comparisons to what young people were doing at the same age 10 years ago, giving a more nuanced picture of how life for young people in England has been changing.

Overall, this is a great resource which offers interested statisticians and researchers exciting possibilities for analyses.

10.9 Invited - Prize winners: Young Statisticians Meeting 2016
Thursday 7 September – 2pm-3pm

Applying quantitative bias analysis to estimate the plausible effects of selection bias in a cluster randomised controlled trial: a secondary analysis

Lauren Barnett, Martyn Lewis, Christian Mallen, George Peat
Keele University

Selection bias is a concern when designing cluster randomised trials (c-RCT). Despite addressing potential issues at the design stage, bias cannot always be eradicated from a trial design. The application of bias analysis presents an important step-forward in evaluating whether trial findings are credible.

The aim of this study is to give an example of the technique to quantify potential selection bias in cluster randomised trials.

This analysis uses data from the Primary Care Osteoarthritis Screening Trial (POST). Quantitative bias analysis is a seldom used technique that can quantify types of bias present in studies. Due to lack of information on the selection probability, probabilistic bias analysis with a range of triangular distributions was also used, applied at 6 month follow-up, post-consultation.

Worse pain outcomes were observed among intervention participants than control participants. Probabilistic bias analysis suggested the observed effect became statistically non-significant if the selection probability ratio was between 1.2-1.4.

The use of probabilistic bias analysis in this c-RCT suggested that worse outcomes observed in the intervention arm could plausibly be attributed to selection bias. A very large degree of selection of bias was needed to mask a beneficial effect of intervention making this interpretation less plausible.

11.1 Invited - Recent advances in pharmacovigilance and pharmacoepidemiology: session in honour of David Finney
Thursday 7 September – 3.10pm-4.30pm

Delivering new evidence and risk prediction tools to support safer use of medicines through real world data analytics

Marion Bennie
University of Strathclyde

Objectives: To present an overview of the national prescribing dataset capability in Scotland, its application in pharmacoepidemiology studies and early use in supporting the development of prediction tools for clinical decision support.

Method: The national prescription information (PIS) system covers all National Health Service (NHS) prescriptions prescribed, dispensed and reimbursed within the community setting, covering in Scotland a total population of 5.3 million residents. Since 2009, the addition of a unique numeric identifier, the Community Health Index (CHI), onto prescriptions permits the study of chronological drug use and heath outcomes, through record linkage to other national datasets, complementing the evidence generated through clinical trials and drug surveillance studies.

Results:

PIS is generated through routine clinical systems in the NHS and hosted within a national information warehouse managed by NHS National Services Scotland (NHS NSS). PIS holds information for over 1.6 billion prescriptions reimbursed in the community from January 1993 to 2014, over 507 million items prescribed and over 344 million items dispensed from 2009 to 2014. Investment in PIS, by the FARR Institute @ Scotland a NHS/academic collaboration, is enabling improved drug coding and transformation of drug dosage instructions into structured fields to support the interpretation and quantification of drug exposure and measurement of patient adherence.

These data are now enabling national studies to be conducted, new knowledge to be generated and health policy and clinical practice to be informed and shaped using real world data analytics. The presentation will illustrate application of these data in: shaping antimicrobial stewardship programs; enabling the delivery and monitoring of interventions to improve high-risk prescribing in primary care; and, measuring the effectiveness of new medicines within routine clinical practice.

Conclusion:

PIS is a valuable new resource that can be used locally, nationally and internationally to promote the safe and effective use of medicines.

11.1 Invited - Recent advances in pharmacovigilance and pharmacoepidemiology: session in honour of David Finney
Thursday 7 September – 3.10pm-4.30pm

WOSCOPS 20 year follow-up data of participants and screenees

Colin McCowan, Ian Ford, Chris Packard, Heather Murray, Robin Young, Kevin Ross
University of Glasgow

The West of Scotland Coronary Prevention Study was a primary prevention trial in 45- to 64-year old men with high low-density lipoprotein cholesterol. A total of 6595 men were randomized to receive pravastatin 40 mg once daily or placebo for an average of 4.9 years. Subsequent linkage to electronic health records of this group and an additional 74,000 men permitted analysis of major incident events over 20 years.

This study examined the 6,595 men who participated in the trial over 20 years to calculate the long term benefits and potential risks of an initial 5 years of statin therapy. We also examined cumulative burden of disease over a 20 year period grouped by baseline cholesterol in approximately 70,000 men categorised by presence or absence of prior heart disease. We then modelled the reduction in cumulative events form different LDL-cholesterol lowering scenarios based on this data.

Statin therapy was associated with a 23% reduction in all cause mortality over the 20-year period, mainly due to reduced cardiovascular deaths. Cumulative hospitalisations were also reduced for any coronary event, for myocardial infarction and for heart failure. There was no evidence suggesting any increase in cancer, non-cardiovascular deaths or hospital admissions for non-cardiovascular reasons.

Over the 20 year follow-up there was a marked increase of over 40 CHD events per 100 subjects for men with no prior history of heart disease in the highest baseline cholesterol group compared to the lowest. Models suggested that lowering cholesterol by 1.0 mmol/l would result in reduction of almost 9 CHD events per 100 subjects for men with no prior heart disease and between 27-36 for those with a history of CHD.

11.1 Invited - Recent advances in pharmacovigilance and pharmacoepidemiology: session in honour of David Finney
Thursday 7 September – 3.10pm-4.30pm

David Finney: A Man for all Seasons

Ivor Ralph Edwards
Uppsala Monitoring Centre

A global database of 15 million individual case harm reports of suspected adverse drug reactions (ICHR) from 150 countries globally has all kinds of useful information, managed by the UMC.

David Finney suggested to WHO that such a collection be made to allow the singular experiences of clinicians suspecting adverse drug reactions be collated, to suggest hypotheses of harm from drugs for further verification.

Over nearly 50 years it has been found serious effects from drug adverse reactions are relatively rare but the overall impact on health care is great. The work of the UMC is as an adjunct to national pharmacovigilance centres throughout the world and revolves around the analysis of 'big data'.

This work involves following reporting trends for disproportional deviations of ICHR reporting from the background norm using a Bayesian approach. Each ICHR is scored for data contained within it that has value in making a causal assessment in that case - a diagnosis. The results are collated and hypotheses made using, inter alia, the Bradford Hill guide-lines.

Since causation of an adverse clinical effect by a drug is usually complicated we have evolved pattern recognition approaches to assess all the surrounding contextual evidence that is available around each case, and this includes work started on text analysis.

Both regulatory and clinical decisions have often to be based on incomplete and heterogeneous data. We have been working on how effectiveness and risk might be considered rationally using ALL data available and accepting that the data on effectiveness of of a drug is soundly based but in limited, selected data from clinical trials, and that harm is often qualitative and from observational studies.

Our aim is to give the best help not only to regulators but also to clinicians and, primarily, their patients.

11.3 Invited - Novel statistical approaches applied to geospatial health
Thursday 7 September – 3.10pm-4.30pm

Bayesian Disease Mapping and Dynamic Surveillance Designs

Annibale Biggeri
University of Florence

Bayesian Disease Mapping is widely performed in Infectious and Parasitological Surveillance. We introduced a tri-level hierarchical Bayesian model to model posterior predictive probabilities of infection (PPP) . In a multivariate framework, PPPs are useful in farm profiling to tailor specific interventions. (Catelan, 2012) This modeling approach can naturally be extend to the design phase of sampling surveys. (Musella, 2014) We aim to introduce any survey in a dynamic surveillance process in which sampling probabilities are continuously updated. We show the connections with Thompson’s adaptive web sampling (2006) and Preferential sampling. (Cecconi, 2016a 2016b).

Cecconi L, Grisotto L, Catelan D, Lagazio C, Berrocal V, Biggeri A. Preferential sampling and Bayesian geostatistics: Statistical modeling and examples. Stat Methods Med Res. 2016 Aug;25(4):1224-43. doi:10.1177/0962280216660409.

Musella V, Rinaldi L, Lagazio C, Cringoli G, Biggeri A, Catelan D. On the use of posterior predictive probabilities and prediction uncertainty to tailor informative sampling for parasitological surveillance in livestock. Vet Parasitol. 2014 Sep 15;205(1-2):158-68. doi: 10.1016/j.vetpar.2014.07.004.

Catelan D, Rinaldi L, Musella V, Cringoli G, Biggeri A. Statistical approaches for farm and parasitic risk profiling in geographical veterinary epidemiology. Stat Methods Med Res. 2012 Oct;21(5):531-43. doi: 10.1177/0962280212446329.

11.3 Invited - Novel statistical approaches applied to geospatial health
Thursday 7 September – 3.10pm-4.30pm

Widening statistical participation in combatting Neglected Tropical Diseases under conditions of global change and uncertainty

Mark Booth
Newcastle University

The term Neglected Tropical Diseases (NTDs) summarises the status of dozens of infectious diseases that collectively affect the health of hundreds of millions of individuals across the globe. They are termed as 'neglected' due to a historic lack of investment in research and control. Current efforts to eliminate these infections have been only partly successful for reasons that are not fully understood. The life-history traits of each species are affected by abiotic and biotic factors acting dependently and independently, resulting in fragmented geographical and demographic distribution. Dynamic and statistical models are available for some species, and have been used to guide control efforts, but no-one modelling approach dominates or captures the complexity in adequate detail, and the abundance of available information relevant to the modelling process has not been comprehensively explored or exploited. There is considerable scope to widen participation amongst statisticians, particularly in the context of developing models that are adaptive to future uncertainties. Here, I describe the outcome of combining statistical and agent-based models to construct spatially explicit models of the future transmission of a parasitic infection in East Africa, under several climate change scenarios. This approach to modelling uncertainty serves as a framework for future efforts to optimise our understanding of how global change could affect regional and local changes in hazards, risks and vulnerabilities associated with NTDs.

11.3 Invited - Novel statistical approaches applied to geospatial health
Thursday 7 September – 3.10pm-4.30pm

Use of spatial analysis in public health: passive and active surveillance to define spatial heterogeneity of infectious diseases

Donal Bisanzio
University of Oxford

Infectious diseases show spatial heterogeneity, which is an important factor to account for when planning surveillance and control systems. Data collected by active and passive surveillance can be used to investigate spatial heterogeneity of infectious diseases. Insights obtained by such analyses are able to highlight hot-spots at different spatial resolutions. However, data collection plans have to take into account the available resources which are usually scarce in developing settings. Co-existence of several diseases (e.g., malaria and soil transmitted helminths) threatens the health of populations living in developing countries. Therefore, identifying targets for control systems is vital for reducing the burden of infectious diseases. Such hot-spot search can be performed by applying spatial analysis procedures using datasets created from merging remote sensing data with data collected from households or health care facilities. Studies performed using these methodologies have shown high capability of identifying areas at high risk of disease transmission at different scale levels. Data collected by active surveillance has the ability to identify hot-spots of co-infections at fine scale which are able to guide control system planning at the household level. However, due to high costs, active surveillance it is not always feasible in areas with few resources and a large territory to survey. Therefore, in such settings, spatial analysis applied to data collected with passive surveillance becomes the best cost-effective procedure for identifying disease hot-spots. Although the spatial level of passive surveillance rarely matches that of active surveillance, it is possible to identify hot-spots at the village level using data collected at healthcare facilities. Thus, passive surveillance can be useful for identifying the high-risk populations and priorities for disease control interventions. Information on spatial heterogeneity of diseases can guide surveillance by targeting specific areas and thus reducing sampling costs for public health institutions.

11.4 Invited - Best of MathSport International
Thursday 7 September – 3.10pm-4.30pm

Detecting Tennis Match-Fixing in In-Play Markets

Oliver Hatfield, Jonathan Tawn, Christopher Kirkbride, Timothy Paulden, David Irons, Grace Stirling
Lancaster University

Match-fixing presents a serious and challenging problem for sport in the 21st century. Tennis has seen particularly widespread coverage of its match-fixing issues in 2016, and it is here that our particular research is concentrated. Although match-fixing is, by design, difficult to detect, the literature demonstrates that the betting activity of match-fixers may lead to distortions in the betting markets. We would therefore expect the predictions of (sufficiently accurate) sports models to show a larger discrepancy with the odds in fixed matches than they would otherwise. Moreover, while the majority of match-fixing detection in the literature considers pre-match markets, our focus is on detecting match-fixing within an in-running
scenario.

A key challenge in assessing the discrepancies between odds data and statistical predictions lies in understanding the uncertainty of both measures. Towards this goal, we extend the widely-used Glicko model to provide estimates of match-win probabilities in tennis with uncertainty, using data from ATP World Tour, Challenger Tour and Futures tournaments from 1991-2016. We then utilise the combinatorial structure of tennis to update our pre-match predictions throughout the match, maintaining uncertainty in a novel manner. Our predictions are compared with live odds data (on a game-by-game level) from 265 men's matches, provided by ATASS Sports, with whom we have collaborated on this project. We also explore uncertainty that arises in the generation of odds across different matches. While our early work shows promising results, further work remains to reliably pick up some of the more subtle anomalies.

11.4 Invited - Best of MathSport International
Thursday 7 September – 3.10pm-4.30pm

The Nappy Factor in Golf: The Effect of Children on the Sporting Performance of Professional Golfers

Tony Syme
University of Salford

This study investigates the effects of children upon the sporting performance of fathers who are professional golfers. Biographical and sporting data for 225 professional golfers are used to estimate fixed-effects regressions. In line with other studies, it is found that performance and earnings improve significantly after the birth of the first child and that this declines after each subsequent child. The fatherhood premium, or ‘nappy factor’, is estimated to be an increase in earnings of 10% for any first child, but this rises to 16% if the first-born child is a son and remains the only child. This study suggests that the rank-order nature of tournaments and the non-linear distribution of prize money within professional golf creates positive incentives to increase work effort, but that tournaments also increase pressure, particularly towards the end of tournament, and that ability to perform under pressure is increased if the player has a son as a first-born child.

11.4 Invited - Best of MathSport International
Thursday 7 September – 3.10pm-4.30pm

On scoring rates and competitive balance in international rugby union

Phil Scarf
University of Salford

We show that scoring rates in international rugby union have more than doubled over the last 50 years. Further, in the double Poisson match, more scoring implies less uncertainty of outcome and therefore less suspense and fewer surprises. We argue that scoring in Rugby approximates to a double Poisson match, so that more scoring in the game is arguably not good for its long-term development. We suggest modifications to the scoring system that would reduce scoring rates and explore these modifications through a rugby world cup tournament simulation.

11.7 Invited - Facilitating research use of administrative data
Thursday 7 September – 3.10pm-4.30pm

Synthetic data in practice: software, applications and challenges

Beata Nowok, Gillian Raab, Chris Dibben
University of Edinburgh, Administrative Data Research Centre - Scotland (ADRC-S)

Confidentiality constraints often restrict access to individual level data. One of the methods that can be used to resolve this problem is the creation of synthetic datasets which will give similar results to those that would be obtained from the real data, but will contain no records that correspond to real individuals. It is achieved by replacing sensitive records with values generated from probability distributions estimated from the original confidential data.

At the Administrative Data Research Centre - Scotland (ADRC-S) we have been developing methodology for generating synthetic data and implementing it in the freely available *synthpop* package for R. The software simplifies considerably the synthesising process and allows staff of statistical agencies to create synthetic datasets that are safe to be released to users. It offers a choice between different parametric and non-parametric synthesising methods. The latter includes classification and regression trees (CART) models. The *synthpop* package provides also tools to assess quality of the synthetic datasets and to apply additional means of confidentiality protection.

This presentation will introduce the *synthpop* package for R and illustrate its functionality with empirical examples. The package is used to produce synthetic data for the users of the Scottish Longitudinal Study. This and other applications of synthetic data, and challenges that still need to be overcome will be discussed.

11.7 Invited - Facilitating research use of administrative data
Thursday 7 September – 3.10pm-4.30pm

ADRN: supporting access to functionally anonymised linked administrative microdata

Elaine Mackey
University of Manchester

The Administrative Data Research Network (ADRN) is a new legal and efficient pathway supporting research access to linked administrative microdata. In this talk I will explain how the ADRN enables researchers to access administrative data that is both detailed and functionally anonymised. By way of doing this, I apply the Anonymisation Decision-making Framework (Elliot, Mackey, O'Hara and Tudor 2016) as a basis for considering the technical, legal, social and ethical aspects of anonymised data within the ADRN Context.

11.7 Invited - Facilitating research use of administrative data
Thursday 7 September – 3.10pm-4.30pm

Synthetic data as a method for protecting confidentiality

Robin Mitra
University of Southampton

Synthetic data is an increasingly popular approach used to protect confidentiality in sensitive data. In this approach, data-holders model relationships between variables in the data and then replace values in the data with synthetic values that are drawn from this statistical model. This data set is then released to analysts. As the released data now contains a certain proportion of synthetic values, confidentiality has been protect to an extent. In addition, provided a plausible model has been used to generate the synthetic values the statistical properties present in the original data should be preserved in the released data.

In this talk we briefly review the synthetic data approach and consider applications to administrative data.

11.8 Invited - Financing companies
Thursday 7 September – 3.10pm-4.30pm

Applying Airline Optimization Modelling to Finance Companies

William Fite
Cahokia Point

This presentation will:

- review airline revenue management modeling describing components of a revenue management system.
- diagrammatically illustrate its applicability to the hospitality sector.
- compare and contrast airline / hospitality network structures with a similar finance company structure.
- discuss the component models necessary to maximize finance company revenues comparing them to airline revenue management systems.
- provide a framework for a holistic model of finance company revenue maximization.

11.8 Invited - Financing companies
Thursday 7 September – 3.10pm-4.30pm

Small datasets – Analyses & Strategies in Corporate Banking

Michelle Greenidge
MLG STATISTICAL SERVICES

In these days of BIG DATA, there is still often a requirement to analyse, draw useful conclusions and build models using small data sets. This is often the case in corporate banking where sample sizes may be relatively small and item/events of interest can be few and indeed sometimes non-existent in the sample.

This talk illustrates some of the issues and suggests some potential strategies using an amalgam of corporate banking Probability of Default model builds to illustrate the points raised. To illustrate some approaches that may be used; an example is taken of an historic portfolio where the number of customers never exceeded 64 at any point, and no default event ever occurred. The requirement was for a probability of default model that would be used to rate individual customers within that portfolio. This was to address the business need for estimating credit risk for a group of customers from whom the bank had the potential to lose approximately £10BN.

The issues on which this talk will touch are:

- The context of such exercises
- Stakeholders – who are they and how are they to managed?
- What do you build on when the dataset is small, and the events of interest are not present?
- How do you validate? Indeed, can you validate?

11.8 Invited - Financing companies
Thursday 7 September – 3.10pm-4.30pm

Simultaneity of Bank Risk and Return

Sean Harkin, Jonathan Crook, Davide Mare
Credit Research Centre, University of Edinburgh Business School

Empirical studies have sought to estimate effects of bank ownership and governance on risk and return, in order to guide reforms to banking. However, such studies have omitted simultaneity between risk and return. In doing so, they neglect the prediction of finance theory that risk and return are simultaneous and may experience bias through confounding of regressor effects with simultaneous effects. For the first time, we show simultaneity between risk and return, measured at the level of bank financial accounting data, and demonstrate that neglect of this simultaneity induces bias in measuring the effects of ownership and governance variables. This is important firstly because it confirms that relationships predicted by finance theory hold in this setting and are not fully obscured by information asymmetries or behavioural biases. Secondly, it provides an unbiased framework in which we analyse effects of governance and ownership, identifying relationships with key implications for policy. Thirdly, in the context of financing banks, it shows investors and regulators that earnings at the bank level are linked to risk. This, in turn, might imply the banking system is transparent enough for regulatory measures aimed at reducing banks' risk to also reduce banks' costs of debt and equity financing.

Plenary 7 - Significance Lecture
Thursday 7 September – 4.50pm-5.50pm

If You Don't need (Astro)Statistics, You Have Done the Wrong Experiment

Roberto Trotta
Imperial College London

At the beginning of last century, the physicist and Nobel Prize Winner Ernest Rutherford reportedly believed that "If your experiment needs statistics, you ought to have done a better experiment". If he were alive today, he probably would not recognize the way cosmology (the study of the Universe on its largest scale) has developed: essentially all of the exciting discoveries in the last two decades have relied on sophisticated statistical analyses of very large and complex datasets. Today, advanced astrostatistical methods belong to the toolbox of almost every cosmologist.

In this talk I will give an overview of how cosmologists have established a "cosmological concordance model" that explains extraordinarily well very accurate observations ranging from the relic radiation from the Big Bang to the distribution of galaxies in the sky in the modern Universe. The emerging picture of a cosmos remains puzzling: 95% of the Universe is constituted of unknown components, dark matter and dark energy. Our understanding of the Universe is -- already today -- limited by our statistical and computational methods. I will discuss how astrostatistics will meet the challenges posed by upcoming extremely large data sets and thereby be instrumental in answering some of the most fundamental questions about the physical reality of the cosmos.

INDEX

ROYAL
STATISTICAL
SOCIETY

DATA | EVIDENCE | DECISIONS