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

1.1 Contributed - Methods & Theory: GLMs and Compositional Data

Tuesday 6 September 2016 – 9am-10am

How to get along with Sum 1: Generalized Linear Models for Compositional Data

David Firth, Fiona Sammut
University of Warwick

Multivariate measurements in the form of a composition, i.e., a vector of proportions or percentages that sum to 1, are of interest in a wide variety of application fields. A classic application area is geology, where rock samples are weighed into their constituent parts (e.g., different minerals). Other examples include elections (percent votes for different parties), geography (percent of land area used for different purposes), and of course statistics (probabilities on a finite sample space). We consider regression analysis, where interest is in the dependence of a composition upon explanatory variables. The most standard approach is from well-known work by J Aitchison: use standard linear models, with multivariate normal assumptions, for a reduced vector of /log-ratios/ of components. Aitchison's approach overcomes the inherent difficulty that the components sum to 1, by analysing only their relative values (on the log scale). In our work reported here, an alternative approach is developed: a natural generalized linear model which avoids data-transformation. This new approach extends the seminal work of Wedderburn (1974, Biometrika): it overcomes the usual difficulties with interpretation of linear models for log-measurements, and difficulties also when there are zeros in the data.

1.1 Contributed - Methods & Theory: GLMs and Compositional Data

Tuesday 6 September 2016 – 9am-10am

High-dimensional Inference for Gamma and Inverse-Gaussian GLMs

Hassan Pazira, Ernst Wit
University of Groningen

The dgLARS method (the method proposed in Augugliaro, Mineo, and Wit, 2013) is based on a differential geometrical extension of the least angle regression method proposed in Efron, Hastie, Johnstone, and Tibshirani (2004). We extend this method for studying the sparse structure of a generalized linear model (GLM) to a larger class of the exponential family, namely the \textit{exponential dispersion family} (when the dispersion parameter, ϕ, is known and unknown). In our algorithm, we use a new developed predictor-corrector (DPC) method such that after the corrector step, the 'optimal' step size from the previous step is reduced using a new method called Regula-Falsi (rf) method, so that it reduces the number of the points of the solution curve. In other words, using the DPC algorithm leads to decrease the number of iterations, and hence our algorithm is faster than the used algorithm in Augugliaro \textit{et al.} (2013 and 2014a) that used the contractor factor (cf). We present our new package that implements the proposed PC algorithm, with both cf and rf method, to compute the solution curve implicitly defined by dgLARS based on Gamma and Inverse Gaussian regression problems with the most commonly used link functions. At the meantime, we develop our package by adding Logistic regression model, as a spacial case, to be able to compare results to the \texttt{dglars} package.

1.1 Contributed - Methods & Theory: GLMs and Compositional Data

Tuesday 6 September 2016 – 9am-10am

A parametric approach to count zeros imputation in compositional data sets

Marc Comas-Cufí, Javier Palarea-Albaladejo, Josep A. Martín-Fernández, Glòria Mateu-Figueras
University of Girona

Compositional data (coda) are multivariate positive observations representing relative contributions to a total. The fact that the data are understood as relative amounts is commonly reflected by expressing them in percentage or similar units. Typical examples include chemical or nutritional compositions, time use budgets and so on. The log-ratio methodology has become the mainstream approach to tackle their statistical analysis. Basically, it assumes the relevant information is contained in the ratios between the parts. However, the use of ratios meets practical difficulties when the observed coda contain zeros. Hence, data pre-processing frequently involves the imputation of zeros by sensible values. A number of proposals have been introduced for the case in which zeros occur in continuous coda as a consequence of a censoring problem.

The equivalent zero problem in the context of discrete compositional count data has been only recently approached from a non-parametric point of view. Compositional count data are discrete vectors accounting for the number of outcomes falling into an array of mutually exclusive categories. A compositional analysis is deemed appropriate when the total sum of those vectors is not of interest. Zero values usually happen in the most infrequent categories as a result of insufficient number of trials.

We propose a parametric count zero replacement based on compounding a log-ratio normal distribution defined on the simplex, the sample space of compositional data, and a multinomial distribution. The log-ratio normal is used to model the vector of probabilities of the categories, whereas the multinomial deals with the counts. Parameter estimation is conducted by maximum likelihood though a Monte Carlo implementation of the Expectation-Maximisation (EM) algorithm.

The estimated multinomial probabilities can be used to impute the count zeros without altering the relative relationships structure of the components. The performance of the method is illustrated by real and simulated data.

1.2 Contributed - Medical Statistics: Surrogacy and Prediction

Tuesday 6 September 2016 – 9am-10am

Toward quantification of the strength of surrogate biomarkers from observational clinical studies

Gabriela Czanner, Ian MacCormick, Simon Harding, Brian Faragher
University of Liverpool

Evaluation of surrogate biomarkers is an important and challenging task and traditionally focussed on evaluating surrogate biomarkers in clinical trials. Other approaches that have been developed do not assume randomisation and include validation metrics such as the likelihood reduction fraction (LRF) and proportion of information gain (PIG). These metrics are based on the intuitive approach of assessing the information in the surrogate about the true endpoint. However they do not evaluate this information within the context of uncertainty, or noise, in the true endpoint. Indeed such uncertainty appears to have received very little attention in the literature on evaluation of surrogate biomarkers.

Our objective is to propose a metric to evaluate the strength of a surrogate biomarker in the presence of noise in the true endpoint variable. We propose a novel metric, called the signal-to-noise ratio (SNR), that evaluates the information in the surrogate marker, relative to the uncertainty in the endpoint. We show how the SNR links with established concepts of statistics (such as Receiver Operating Curve for binary endpoints) and information theory (such as Kullback-Leibler divergence), and we show how it complements the measures of LRF and PIG.

We apply our approach to simulated and to observational clinical datasets from diabetic and malarial retinopathy in order to evaluate the strength of potential surrogate biomarkers.

In this talk we demonstrate that (i) SNR is an important metric to assess the amount of unconditional information in surrogate marker about outcome relative to the uncertainty in the outcome; (ii) we discuss the three metrics SNR, LRF and PIG - each of them evaluates different aspects of the surrogacy; and (iii) we discuss how a combination of biomarkers can be selected and evaluated for their surrogacy.

1.2 Contributed - Medical Statistics: Surrogacy and Prediction

Tuesday 6 September 2016 – 9am-10am

Clinical Prediction in Defined Populations: when and how to aggregate existing models

Glen Martin, Mamas Mamas, Niels Peek, Iain Buchan, Matthew Sperrin
University of Manchester

Background

Clinical prediction models (CPMs) are increasingly deployed to support local healthcare decisions. However, local populations or providers may have too few data for deriving a CPM. Consequently, CPMs previously developed for similar outcomes and populations can be exploited. An emerging approach is to aggregate and recalibrate existing models to the target population. This simulation study aimed to investigated the impact of between-population-heterogeneity and local population sample size on aggregating existing CPMs, compared with developing new ones, in terms of performance measures and mean square error.

Methods

The simulation process generated multiple logistic regression models representing the existing CPMs, where each was fitted to a non-overlapping subset of observations and included a potentially overlapping subset of risk predictors. A further, independent, sample of observations was used to develop a new logistic regression model using backwards selection and ridge regression in addition to aggregating the existing CPMs; aggregation methods included stacked regression, principal component analysis and partial least squares regression. Heterogeneity between all population was induced by applying random effects to a common generating model across all populations.

Results

When the size of the development data was <10% of the largest dataset used to derive existing CPMs and the standard deviation of the random effects was ≤0.25, aggregation gave better performance than deriving a model independent of prior information. Derivation of new CPMs only gave optimal performance when there were more data available than used to derive existing models, or when there was an unusual local context of CPM use.

Conclusion

In conclusion, this study demonstrates a pragmatic approach to contextualising CPMs to local populations, and advises on modelling strategies. Specifically, aggregation of existing CPMs is beneficial in the development of a new CPM for local healthcare predictions.

1.2 Contributed - Medical Statistics: Surrogacy and Prediction

Tuesday 6 September 2016 – 9am-10am

Using multiple biomarkers to inform personalised treatment recommendations in randomised trials

Matthias Pierce, Richard Emsley, Graham Dunn
University of Manchester

A key strand of the stratified medicine paradigm is the attempt to move beyond a 'one size fits all' approach that recommends treatment based on a population response, towards improving patient outcomes through personalised treatment recommendations (PTRs). A PTR maps a set of biomarkers, predictive of differential treatment response, to a decision of whether to treat or not. An optimal PTR both maximises patient outcomes and is applicable to new subjects. In some situations, multiple biomarkers could be better at establishing PTRs than single biomarkers.

Using data from randomised trials and assuming that the expected outcome can be predicted using a linear combination of treatment, plus the predictive and prognostic effect of biomarkers, a PTR can be constructed using a weighted sum of predictive biomarkers and treatment effect. A common approach for estimating weights in this context is to fit a regression model to data, with interaction terms. An alternative method for estimating these weights was proposed by Kraemer and exploits the fact that in this situation we are only interested in the predictive and not the prognostic effect of markers.

We describe the regression and the Kraemer methods for combining multiple predictive makers, and use Monte Carlo simulations to compare the approaches under various data generation scenarios. The simulations utilise a parameter for quantifying the expected benefit of a treatment recommendation compared to a treatment-as-usual approach. These methods for constructing a PTR are contrasted with classification methods that do not rely on the assumption of linearity.

1.3 Contributed - Social Statistics: Missing Data

Tuesday 6 September 2016 – 9am-10am

How to impute missing confounders when using propensity scores

Jonathan Bartlett
AstraZeneca

Propensity scores are increasingly used in the analysis of observational data from economic, social and medical settings to adjust for measured confounders. A common issue in practice is that one or more confounders are be missing for some individuals, complicating analysis. Multiple imputation has previously been proposed to handle such missingness, with missing confounders imputed using exposure status and other confounders, but ignoring outcome. Through a simple example we demonstrate that this approach leads to invalid inferences in general. Instead, we advocate and present results for an approach which multiply imputes missing confounders conditional on outcome, separately in the exposed and unexposed individuals.

1.3 Contributed - Social Statistics: Missing Data

Tuesday 6 September 2016 – 9am-10am

Multilevel modelling approach to analysing socioeconomic status longitudinal data and compensating for missingness

Adrian Byrne, Mark Tranmer, Natalie Shlomo
University of Manchester

Multilevel modelling offers a unique framework for analysing longitudinal data as the method accounts for correlations of observations across time. These data can consist of repeated observations over time (level 1) nested within individuals (level 2). This framework can handle a variety of functional forms of change over time and can tolerate both unequally spaced data and missing data that is assumed to be missing at random.

Missing data may cause bias in parameter estimation and weaken the generalizability of the results in longitudinal studies. Moreover, ignoring cases with missing data may lead to a loss of information which in turn decreases statistical power. Multiple imputation provides one solution to compensate for missing data by replacing missing values with model-based predictions thereby enhancing statistical power and possibly producing less biased model results.

This paper examines the changes in socioeconomic status over the life course by comparing two models: a polynomial growth multilevel model and a step function growth multilevel model both dependent on gender, region of residence and parental socioeconomic status. We also address the problem of missing data, compare multiple imputation solutions and contrast these with complete-case and available-case results.

The Occupational Earnings Scale is chosen as the measure of socioeconomic status (Bukodi, Dex and Goldthorpe 2010; Nickell 1982). This measure injects a form of hierarchy into routinely collected occupation data by ordering each occupation according to its mean hourly wage rate (ONS Annual Survey of Hours and Earnings) and produces a continuous measure of socioeconomic status over the life course. The 1958 National Child Development Study forms the longitudinal dataset (University of London, Institute of Education, Centre for Longitudinal Studies) and has the property that data collection is not evenly spaced across the life course.

1.3 Contributed - Social Statistics: Missing Data

Tuesday 6 September 2016 – 9am-10am

Handling missing data models in the presence of weighting

Harvey Goldstein, James Carpenter, Michael Kenward
University of Bristol

Multiple imputation (MI) for missing data is now a well-documented procedure implemented in a variety of software packages (Carpenter and Kenward, 2013). There is, however, little discussion of the case where weights are assigned to data records as often occurs in sample surveys. While algorithms based upon chained equations such as MICE (White, Royston and Wood, 2011) are able to incorporate such weights at both the imputation stage and when fitting the model of interest, joint estimation algorithms do not readily incorporate these. We propose a new hybrid frequentist/Bayesian approach that uses a 2-stage bootstrap wrapped around a Bayesian imputation model, and leads to consistent estimators. Specifically, we extend the Bayesian model for missing data proposed by Goldstein, Carpenter and Browne (2014) which extended existing methods such as chained equations, by allowing for interactions among variables having missing values.

The talk will describe the methodology and provide results from simulations and a substantive application using data from the Millenium Cohort Study.

References:

White, IR, Royston, P, Wood, AM (2011). Multiple imputation using chained equations: Issues and guidance for practice. Statistics in Medicine*,* 30*, 377–399.*

Carpenter JR; Kenward MG. (2013). Multiple Imputation and its Application*. John Wiley & Sons Ltd: Chichester.*

Goldstein, H., Carpenter, J. R. and Browne, W. J. (2014), Fitting multilevel multivariate models with missing data in responses and covariates that may include interactions and non-linear terms. Journal of the Royal Statistical Society: Series A (Statistics in Society). 177(2), 553-564
doi: 10.1111/rssa.12022

1.6 Contributed - Official Statistics: Productivity & Tax

Tuesday 6 September 2016 – 9am-10am

Seeing the trees for the wood: a cohort approach to tracking the evolution of productivity

Michael Anyadike-Danes
Aston Business School and Enterprise Research Centre

It is commonly presumed that as more data becomes available in a field of study the subject becomes better understood. This does not seem to have been the case with productivity. For almost two decades researchers in the UK have had access to firm-level data covering the population of firms but little progress seems to be made in understanding productivity since the early 2000s. The most striking finding from the early work was the extraordinary degree of dispersion in productivity levels. For example, with a detailed (80 sector) industrial classification a typical finding was that only 15% of dispersion was attributable to inter-industry variation, the other 85% is between firms within industries. In other words firm-level heterogeneity is a hugely significant factor, and it is this heterogeneity which seems to have deterred further work. Here we 'tame' heterogeneity to some extent, but without entirely suppressing the rich diversity of firm performance, by focusing on a single birth cohort of firms and study the evolution of their productivity over a relatively long period of time. Specifically, we follow the 1998 birth cohort up to age 15, an analytical strategy which has proved successful in building an improved understanding of firm-level job growth. Our data record the labour productivity of the 25,000 firms from the birth cohort which survived to 2013 (by age 15 the other 90% of the 250,000 cohort members were dead). Two groups of high performing firms stand out amongst these survivors: the top 5% of the productivity growth distribution record an 30% loss of jobs; whilst the top 5% of the job growth distribution record a productivity decline of about 50%. Whilst these are truly exceptional firms, this finding serves to highlight a difficult choice facing any government committed to both raising employment and championing improved productivity.

1.6 Contributed - Official Statistics: Productivity & Tax

Tuesday 6 September 2016 – 9am-10am

Far from equilibrium: Wealth reallocation in the United States

Alexander Adamou
London Mathematical Laboratory

Studies of wealth inequality often assume that an observed wealth distribution reflects a system in equilibrium. This constraint is rarely tested empirically. We introduce a simple model that allows equilibrium but does not assume it. To geometric Brownian motion (GBM) we add reallocation: all individuals contribute in proportion to their wealth and receive equal shares of the amount collected. We fit the reallocation rate parameter required for the model to reproduce observed wealth inequality in the United States from 1917 to 2012. We find that this rate was positive until the 1980s, after which it became negative and of increasing magnitude. With negative reallocation, the system cannot equilibrate. Even with the positive reallocation rates observed, equilibration is too slow to be practically relevant. Therefore, studies which assume equilibrium must be treated skeptically. By design they are unable to detect the dramatic conditions found here when data are analysed without this constraint.

Joint work with Yonatan Berman (Tel-Aviv University) and Ole Peters (London Mathematical Laboratory, Sante Fe Institute).

1.6 Contributed - Official Statistics: Productivity & Tax

Tuesday 6 September 2016 – 9am-10am

Productivity statistics and the economy

Geoff Tily
Trades Union Congress

The presentation will offer alternative perspectives on productivity statistics, drawing on UK and OECD National Accounts and labour market statistics.

Outcomes are given great prominence as indicative of supply failures with the UK economy. A demand interpretation will be presented. Background figures on the cylicality of headline outcomes and parallel movements across all countries will be shown. Aggregate national accounts expenditure and income figures (for the UK and OECD) will be decomposed first to show aggregate GDP outcomes are driven by demand. The income figures, supplemented with employment and earnings statistics, will then show how the labour market has adjusted to weak growth through reduced price (i.e. earnings growth) rather than quantity (i.e. employment). Productivity is then a residual of these processes, rather than factor with causal force. A third analysis will decompose productivity outcomes across the OECD bewteen growth and employment, to show that productivity can abstract from underlying differences of importance (and eg show how high productivity in Spain is a worse outcome than the lower productivity in the UK). The policy implications are of the greatest importance, given policmakers continue to interpret headline figures as proving supply flaws with the economy. If demand is the dominant factor (at least in the short run), policy inaction may lead to deflation. The work will be underpinned by analysis already in the public domain, but updated and extended. See here:
https://www.tuc.org.uk/sites/default/files/productivitypuzzle.pdf

1.7 Contributed - Industry & Commerce: Evaluating Customers and Supply Chains

Tuesday 6 September 2016 – 9am-10am

ECrfBimax Algoritm: Using Bicluster Analysis to Improve Reference Class Forecasting

Gloria Gheno, Massimo Garbuio
Ca' Foscari University, ECLT

Background

Recent psychological studies as well as the planning community have championed reference class forecasting (RCF) as a superior alternative to statistical modeling. However, there is only scattered evidence of the value of RCF in managerial decision making.

Objectives

We propose a variation to RCF by using a bi-clustering approach, applied to a set of qualitative data. The data identify the attributes of products and services. Our algorithm, which we call ECrfBimax, identifies the extent to which companies are targeting clients using similar as well as dissimilar offerings.

Methods

The bi-clustering approach is used to extrapolate reference classes from a proprietary, high-dimensional dataset comprised of 10 companies and a set of 200 clients. We estimated similarities in offerings across a range of operating companies. The robustness of the bi-clusters is measured by considering quantitative measures of company performance, controlling for the age of the company and industry.

Conclusion

The resulting bi-cluster groups the clients and the characteristics, which are or are not developed by companies in a hierarchical way. Indeed, the algorithms proposed in the literature for binary variables, where the event of interest is codified equal to 1, do not provide the possibility of understanding the importance of the variables and of analyzing their conditional probabilities directly from the obtained bi-clusters. Our algorithm instead produces bi-clusters where it is possible to identify the characteristics that are never developed under the condition of the development of other characteristics grouping in a second step the zero elements, which are not considered in the traditional bi-cluster literature. The procedure allows us to isolate an offering’s most important set of characteristics.

From a managerial perspective, our method provides meaningful information to identify which product attributes are appealing to different customer segments as well as the most urgent to be addressed.

1.7 Contributed - Industry & Commerce: Evaluating Customers and Supply Chains

Tuesday 6 September 2016 – 9am-10am

Developing a Multivariate Bullwhip Effect Measure

Chaitra Nagaraja, Tucker McElroy
Fordham University

The bullwhip phenomenon exists when the variability of the quantity ordered outpaces the variability of demand. When this effect is present, it is difficult to maintain control of inventory levels. Furthermore, retailers must manage supply chains across multiple suppliers and demand for their products can be interdependent. To incorporate these practical features, we consider multivariate demand models for which the differenced stationary vector time series has a Wold representation and where general forecasting formulas are available. Using this general class of models, we derive a multivariate bullwhip effect measure for a two-stage supply chain with an order-up-to inventory policy. This corresponds to one retailer handling *m*, possibly interdependent, products. Orders are placed simultaneously with a fixed, common lead-time horizon (i.e., time between placing and receiving an order). We illustrate our method and its implications on supply chain management using the Dominick's Database, published by the James M. Kilts Center, University of Chicago Booth School of Business.

1.7 Contributed - Industry & Commerce: Evaluating Customers and Supply Chains

Tuesday 6 September 2016 – 9am-10am

Rapid Quantification of Specification Risk for Scorecards

Alan Forrest
Royal Bank of Scotland

Scorecard models, usually based on logistic regression, are used by Banks and Financial Institutions to evaluate Borrowers' Credit Risk. The validity of these models is critical to each Bank's operations and profitability, but the industry recognises that drift in customer population, background economy, or in more obscure factors, degrade the models' performance. The models also make assumptions, such as independence of certain components that are open to challenge and change. Banks' Model Risk managers assess and quantify this risk of model failure, and discourage models whose specification is sensitive to data shift and changes in assumption.

This talk presents and illustrates a quick and practical calculation that quantifies this "specification risk" for scorecards. It exploits a view of model selection that mixes geometric and information-theoretic ideas, closely related to developments by Amari and co-authors.

A scorecard on a population, with discrete factors and outcomes, is viewed as an ideal distribution of cell-counts on a contingency table, to be compared with the development data cell-counts: data and model live in the same high dimensional space of cell-counts. Candidate models are constrained to a subspace and the chosen model is the point in model space "closest" to the data. For Maximum Likelihood Estimation the Kullback-Leibler divergence is the natural measure of closeness, albeit asymmetric.

This talk presents an intuitive inequality that constrains the KL-divergence between the original model and its shift, knowing the data shift and the geometric curvature of the model space. This divergence limit to model change is easy to calculate in real applications, and allows managers to quantify specification risk quickly.

Also, the curvature term implies that to minimise specification risk, developers should use the flattest model spaces: in practice, this suggests they should prefer dummy marginal factors and avoid dimension reduction, such as classic weights-of-evidence transformation.

PLENARY 1

Tuesday 6 September 2016 – 10.10am-11.10am

Statistical Paradises and Paradoxes in Big Data

Xiao-Li Meng
Harvard University

Statisticians are increasingly posed seemingly paradoxical questions, challenging our qualifications for entering the statistical paradises created by Big Data. Two such questions represent the use of Big Data for population inferences and individualized predictions: (1) “Which one should I trust: a 1% survey with 60% response rate or a self-reported administrative dataset covering 80% of the population?” and (2) “Personalized treatments -- that sounds heavenly, but where on earth did they find the right guinea pig for me?” Investigating the first question reveals a Big Data Paradox: the bigger the data, the more certain we will miss our target. We need data-quality indexes, not merely quantitative sizes, to answer the question: a seemingly tiny self-reporting bias will make an on-line database with 160,000,000 entries equivalent to a simple random sample of 400 for estimating population averages. The second question is fundamentally yoked to the familiar Simpson's Paradox: how do we ensure that the level of aggregation (i.e., data resolution) does not alter our (treatment) conclusions? A multi-resolution framework, inspired by wavelets, provides a theoretical platform for studying statistical evidence for predicting individual outcomes. In contrast to the first question, where the goal is to infer population quantities from samples, with the second question we seek a primary inference resolution, that is, a sensible bias-variance tradeoff to form populations for approximating individuals. Theoretical links between optimal resolution and sparsity will be discussed.

2.1 INVITED - Methods & Theory: Recent Advances in time/space modelling

Tuesday 6 September 2016 – 11.40am-1pm

Kriging Over Space and Time Based on a Latent Reduced Rank Structure

Qiwei Yao, Da Huang, Rongmao Zhang
London School of Economics and Political Science

We propose a new approach to extract nonparametrically covariance structure of a spatio-temporal process in terms of latent common factors. Though it is formally similar to the existing reduced rank approximation methods (Section 7.1.3 of Cressie and Wikle, 2011), the fundamental difference is that the low-dimensional structure is completely unknown in our setting, which is learned from the data collected irregularly over space but regularly in time. We do not impose any stationarity conditions over space either, as the learning is facilitated by the stationarity in time. Krigings over space and time are carried out based on the learned low-dimensional structure. Their performance is further improved by a newly proposed aggregation method via randomly partitioning the observations accordinly to their locations. A low-dimensional correlation structure also makes the kriging methods scalable to the cases when the data are taken over a large number of locations and/or over a long time period. Asymptotic properties of the proposed methods are established. Illustration with both simulated and real data sets is also reported.

2.1 INVITED - Methods & Theory: Recent Advances in time/space modelling

Tuesday 6 September 2016 – 11.40am-1pm

Discrete Longitudinal Data Modelling with a Mean-Correlation Regression Approach

Cheng Yong Tang, Weiping Zhang, Chenlei Leng
Temple University

Joint mean-covariance regression modelling with unconstrained parameterization has provided statisticians and practitioners a powerful analytical device for characterising covariations between continuous longitudinal responses. How to develop a delineation of such an unconstrained regression framework amongst categorical or discrete longitudinal responses, however, remains an open and challenging problem. This paper studies, for the first time, a novel mean-correlation regression for a family of generic discrete responses. Targeting at the joint distributions of the discrete longitudinal responses, our regression approach is constructed by using an innovative copula model whose correlation parameters are represented by unconstrained hyperspherical coordinates. To overcome the computational intractability in maximising the full likelihood of the discrete responses in practice, we develop a computationally efficient pairwise likelihood approach for estimation. We show that the resulting estimators of the proposed approaches are consistent and asymptotically normal. A pairwise likelihood ratio test is further proposed for statistical inference. We demonstrate the effectiveness, parsimoniousness and desirable performance of the proposed approach by analysing three discrete longitudinal data sets and conducting extensive simulations.

2.2 INVITED - Medical Statistics: Uncertainty vs Utility: the statistical fulcrum of health informatics

Tuesday 6 September 2016 – 11.40am-1pm

High resolution data in mental health: the good, the bad and the ugly.

Maximilian Kerz
King's College London

The potential of apps, smartphones and wearables as tools for predictive, precision and patient-centred medicine is enormous, but as yet unrealised. Application of the latter has allowed us to collect ever-increasing data sets, yet their size brings unexpected challenges in extracting valuable information, detecting novelty and making reliable predictions.

SleepSight is a novel and cost-effective approach to passively monitor wellbeing in individuals with psychosis, living in the community. While the primary focus of the study was to test usability and technological feasibility, it has also enabled us to generate a multidimensional, high resolution dataset from various sensors within users' wearable devices and smartphones.

This talk will discuss the system we deployed to reliably collect high resolution data as well as the challenges and pitfalls when employing various methodologies for extracting valuable information from high-resolution time-series.

2.2 INVITED - Medical Statistics: Uncertainty vs Utility: the statistical fulcrum of health informatics

Tuesday 6 September 2016 – 11.40am-1pm

Analysing Primary Care Databases: a synergy of health informatics and traditional statistics

Evangelos Kontopantelis
University of Manchester

Research with structured Electronic Health Records (EHRs) is expanding as data becomes more accessible, analytic methods advance and the scientific validity of such studies is increasingly accepted. However, data science methodology to enable the rapid searching, extraction, cleaning and analysis of these large, often complex, datasets is less well developed. Preparing a research-ready dataset from EHRs is a complex and time consuming task requiring substantial data science skills, even for simple designs.

The UK has been leading on the development of such repositories, particularly Primary Care Databases (PCDs), with several large and many smaller databases currently in use. This can be attributed to two UK-specific conditions that have favoured such a development. First, the umbrella of a single National Health Service (NHS), using broadly uniform health care procedures across providers. Second, the near-universal adoption by general practices of clinical computer systems with defined interoperability specifications.

Focusing on primary care databases, like the Clinical Practice Research Datalink (CPRD), we will discuss their characteristics, availability and data structure. The advantages and disadvantages of using these resources for heath research will be discussed and relevant methodological and software tools will be highlighted. Examples will also be provided, focusing on research questions where randomised controlled trials are unfeasible.

2.2 INVITED - Medical Statistics: Uncertainty vs Utility: the statistical fulcrum of health informatics

Tuesday 6 September 2016 – 11.40am-1pm

Uncertainty vs. Utility: the statistical fulcrum of Health Informatics

Iain Buchan
The University of Manchester

Health informatics describes the technology and infrastructure needed for the right information to reach the right person at the right time to support the right decision. In an age of big data, this is a rapidly shifting and developing landscape. One new development is the 'Learning Health City': a paradigm in which health data are directly fed into the research pipeline, with this research supporting rapid improvements in healthcare delivery, i.e. with minimal 'data action latency'. However, health data arise from multiple sources and through complex data generating processes, so converting these data into information that can support appropriate clinical decisions represents a substantial statistical challenge. This talk will describe this emerging landscape, particularly the challenges, opportunities, and above all, need, for statistical methods and expertise to support health informatics.

2.2 INVITED - Medical Statistics: Uncertainty vs Utility: the statistical fulcrum of health informatics

Tuesday 6 September 2016 – 11.40am-1pm

Misinformation vs. information: the statistical modelling of health data

Damon Berridge, Robert Crouchley, Daniel Grose
Swansea University Medical School

A statistical model embodies a set of assumptions concerning the generation of the observed data. To quote Box & Draper (1987), 'essentially, all models are wrong, but some are useful'. What makes one model more useful than another may depend on how well a model approximates one's notion of the data generating process; in other words, how realistic are those assumptions.

In this paper, we discuss a number of assumptions that are often made when modelling data. These assumptions include the lack of endogeneity of regressors and the absence of omitted effects. Ignoring such features can result in biased statistical models whose results can misinform the data analyst.

We will illustrate some of the issues related to the confounding of omitted variables and endogeneity by way of an example which uses data on health service use and health insurance taken from the Australian Health Survey (Cameron & Trivedi, 1988).

We use a trivariate Poisson model to relate three response variables:

- Number of prescription drugs taken (PRESCRIB)
- Number of non-prescription drugs taken (NONPRESC)
- Number of visits to the doctor (NVISITS)

to type of health insurance, whilst controlling for a set of secondary explanatory variables including a range of demographics and measures of health status such as number of illnesses in the last two weeks (ILLNESS).

We demonstrate how the roles of PRESCRIB and ILLNESS in the model for NONPRESC change depending on whether or not we account for omitted variables. We also show how the variance-covariance structure of the trivariate Poisson model varies depending on whether or not PRESCRIB is included as a regressor in the model for NONPRESC.

2.4 INVITED - Data Science: Statistical and Computational Challenges in Data Science

Tuesday 6 September 2016 – 11.40am-1pm

Data Science: Where Computation and Statistics Meet?

Neil Lawrence
University of Sheffield

What is data science? A new name for something old perhaps. Nevertheless there is something new happening. Data is being acquired in ways that could never have been envisaged 100 years ago. This is presenting new challenges, and ones that no single field is equipped to face. In this talk we will focus on three separate challenges for data science: 1. Paradoxes of the Data Society, 2. Quantifying the Value of Data, 3. Privacy, loss of control, marginalization. Each of these challenges has particular implications for data science and the interface between computation and statistics. By addressing these challenges now we can ensure that the pitfalls of the data driven society are overcome allowing to reap the benefits.

2.5 INVITED - Spatial Statistics: Identifying spatio-temporal trends & clusters in disease risks

Tuesday 6 September 2016 – 11.40am-1pm

Challenges in understanding the spatio-temporal epidemiology of norovirus infection in England using routine public health surveillance data

Helen Clough, Joanne Hardstaff, John Harris, Sarah O'Brien
University of Liverpool

Noroviruses are the commonest cause of acute gastroenteritis worldwide. Commonly termed "the winter vomiting bug", infections exhibit a highly seasonal trend with most occurring during winter months, although they can occur at any time of the year. Illness is self-limiting but can have serious consequences in some groups. Norovirus-associated costs to patients and the health service in the UK have been estimated at between £63 and £106 million annually (Tam and O'Brien, 2016).

Understanding norovirus epidemiology using routine surveillance data is complicated by several factors, for example:

- Norovirus is not a notifiable disease;
- Reporting practice varies between public health regions;
- Cases from large outbreaks, often in closed settings, are more likely to feature in national reports, resulting in certain age groups being over-represented;
- Patients with sudden onset of diarrhoea and vomiting are discouraged from visiting their GP, reducing the likelihood of laboratory confirmation of infection;
- The sensitivity of tests for norovirus has improved over the time period of interest;
- Aggregated data present specific challenges for spatial analysis.

We describe exploratory analysis of regional weekly reports of norovirus cases collected by Public Health England between 2003 and 2014. We consider spatio-temporal models as described by Meyer, Held and Hohle (2016), which allow estimation of epidemic and endemic components in the data. These preliminary analyses provide insight into variability in reporting practice by region. Interest, however, ultimately concerns spatio-temporal variation in norovirus illness, having allowed for variability in reporting. We discuss challenges for inference posed by sampling biases as outlined. We investigate how random effects models might be used to allow for variable reporting rates between regions. Finally, we consider how data from periods of high diarrhoeal awareness might be used to reduce uncertainty around the true national illness and infection burden.

We gratefully acknowledge Public Health England for supplying the data.

2.5 INVITED - Spatial Statistics: Identifying spatio-temporal trends & clusters in disease risks

Tuesday 6 September 2016 – 11.40am-1pm

Some Challenges in the Analysis of Spatially Referenced Survival Data (and Some Possible Solutions)

Benjamin Taylor, Ziyu Zheng, Barry Rowlingson
Lancaster University

In this talk, we discuss two challenges in the analysis of spatially referenced survival data. The first challenge relates to when the exact location of individuals in space is unknown, but it is known that they reside in a region; such data are common in spatial survival analyses due to confidentiality issues (e.g. SEER cancer data). In this talk we will discuss an alternative to the commonly applied CAR model: by treating the data as though they have arisen from a marked point process, we can deliver effectively continuous inference using inferential machinary developed for aggregated log-Gaussian Cox processes.

The second challenge relates to population-based cancer registry data, which is usually collected over an extended period of time e.g. decades. Individuals enter these registries when they are diagnosed with a cancer and their survival prognoses will depend on the effectiveness of available treatment regimens at the time of entry and also the rapidity with which the cancer was detected. More generally, improvements in survival prognosis can be seen not only as a result of treatment quality or public health awareness but also as a result of biases associated with screening: overdiagnosis, lead time and length biases. In many studies, interest focuses on survival duration and the time of entry into the study is eliminated. In this poster, we introduce a Bayesian spatiotemporal model that captures both entry-time effects as they evolve over the study period and also individual survival duration.

2.5 INVITED - Spatial Statistics: Identifying spatio-temporal trends & clusters in disease risks

Tuesday 6 September 2016 – 11.40am-1pm

A general methodological framework for identifying disease risk spatial clusters based upon mixtures of temporal trends

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

We present a novel general Bayesian hierarchical 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, changepoint, concave and convex 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. The effectiveness of (MC)^3 is demonstrated in a simulation study, before applying the model to hospital admission rates due to respiratory disease in the city of Glasgow between 2002 and 2011. Software for implementing this model will be made freely available as part of the R package CARBayesST.

3.1 Contributed - Methods & Theory: Statistical Mathematics

Tuesday 6 September 2016 – 2pm-3pm

Random Projection Ensemble Classification

Timothy Cannings, Richard Samworth
Statistical Laboratory, University of Cambridge

We introduce a very general method for high-dimensional classification, based on careful combination of the results of applying an arbitrary base classifier to random projections of the feature vectors into a lower-dimensional space. In one special case presented here, the random projections are divided into non-overlapping blocks, and within each block we select the projection yielding the smallest estimate of the test error. Our random projection ensemble classifier then aggregates the results of applying the base classifier on the selected projections, with a data-driven voting threshold to determine the final assignment. Our theoretical results elucidate the effect on performance of increasing the number of projections. Moreover, under a boundary condition implied by the sufficient dimension reduction assumption, we control the test excess risk of the random projection ensemble classifier. A simulation comparison with several other popular high-dimensional classifiers reveals its excellent finite-sample performance.

3.1 Contributed - Methods & Theory: Statistical Mathematics

Tuesday 6 September 2016 – 2pm-3pm

The Elliptically Symmetric Angular Gaussian distribution: a new distribution for modelling data on the sphere

Simon Preston, Phillip Paine, Michail Tsagris, Andrew Wood
University of Nottingham

The angular Gaussian (AG) distribution on the sphere S^{d-1} is analogous to the Fisher-Bingham (FB) distribution on S^{d-1}. The AG distribution is the marginal distribution of the directional component of a multivariate Gaussian vector, X, whereas the FB distribution is the conditional distribution of X given |X| = 1. For typical applications the general AG and FB contain too many parameters to be well estimated from data, and for this reason it is useful to work instead with subfamilies of AG and FB. An important subfamily of FB is the Kent (1982) distribution. The Kent has elliptical contours and its number of free parameters equals that of the multivariate Gaussian in the tangent space to the sphere, a balance of flexibility and complexity ideal for many applications. To our knowledge no analogous subfamily of AG has yet been studied. In this talk we present such a subfamily, which we call the elliptically symmetric angular Gaussian (ESAG) distribution. Advantages of ESAG are that it is extremely easy to simulate from and that it has a density that is quick to evaluate, free of any intractable normalising constants. We demonstrate the use of ESAG for modelling and hypothesis testing in both IID and regression settings.

3.1 Contributed - Methods & Theory: Statistical Mathematics

Tuesday 6 September 2016 – 2pm-3pm

A multiresolution analysis of the rate of a point process

Youssef Taleb, Edward Cohen
Imperial College London

In this work we address the non-parametric estimation of the rate (first-order intensity) of an orderly point process on the real line using a multiresolution wavelet expansion approach. Estimating the rate of a point process is a task of great importance in the understanding of its nature and a wavelet approach lends itself to an interesting time-scale analysis. Implementing Haar wavelets, we find that in the case of a Poisson process the piecewise constant wavelet estimator of the rate has a scaled Poisson distribution. We apply this result in the design of a likelihood ratio test for a multiresolution formulation of the homogeneity of a Poisson process. This formulation is defined through the ideas of J-th level homogeneity and inhomogeneity. We demonstrate this method with simulations and provide Type 1 error and empirical power plots under piecewise triangular rate models.

In order to exploit this multiresolution setup, we also look at the information that the detail spaces are able to provide in developing a deeper understanding of the point process' first order behaviour at different levels of resolution. From a sequential examination of the detail spaces derived from the multiresolution analysis, we demonstrate with further likelihood ratio tests that we can discover homogeneous or inhomogeneous behaviours in different bands of resolutions. We validate these tests with simulations and apply it to daily router data on a large network.

3.2 Contributed - Medical Statistics: Applications of Statistics

Tuesday 6 September 2016 – 2pm-3pm

Data linkage and statistical modelling to provide stratified risk assessment for HAI

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

The use of "real-time" data to support individual patient management could be delivered by building robust risk assessment tools onto existing linked data held by NHS Scotland's Infection Intelligence Platform. 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.

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 prior hospitalisations, care-home residence, comorbidities, exposure to gastric acid suppressive drugs and antibiotic exposure, defined as cumulative (defined daily dose (DDD)) and temporal exposure in the previous 6 months, to the risk of CDI acquisition. Antimicrobial exposure was considered overall and for higher risk broad spectrum antibiotics (4Cs). Associations are assessed using conditional logistic regression and cross-validation used to assess the ability to accurately predict CDI infection. Risk scores are estimated by combining these predictions with age and gender population incidence.

In the period 2010-2013 there were 1446 cases of CDI with 7964 matched controls. A dose-response relationship for exposure to any antimicrobial and, with elevated risk, to the 4C group (1-7 DDDs OR=3.8 rising to OR=17.9 for 29+ DDDs) was found. Risk following exposure to 4C antimicrobials is elevated up to 6 months later (4C OR=12.4 within 1 month, OR=2.6 4-6 months later). The risk was also increased with co-morbidities, previous hospitalisations, care-home residency, increased number of prescriptions, and gastric acid suppression.

Despite limitations to current application in practice,(paucity of in-hospital prescribing data and timeliness constraints), when fully developed this system will enable risk classification to identify patients most at risk of HAI and adverse outcomes to aid clinical decision making.

3.2 Contributed - Medical Statistics: Applications of Statistics

Tuesday 6 September 2016 – 2pm-3pm

Endogenous estimation of hospital waiting list dynamics

Daniela Bond-Smith
Centre for Genetic Origins of Health and Disease, University of Western Australia, and Sydney Children's Hospitals Network

The key challenge in predicting patient waiting times for hospital admission is estimating the dynamic effect of changes in the composition of the waiting list, both from other patients waiting concurrently and from the path-dependent effect of historical changes. A clear shortcoming of previous research is that the proposed models are unable to directly capture the dynamic feedback effects characteristic for hospital waiting lists, because cross-sectional independence is assumed. This is not plausible for modelling waiting lists; in fact we have a substantive interest in explicitly determining the nature of this interdependence and its influence on the predictive value of other determinants of waiting time.

We propose a model of patient waiting time based on an original application of spatial econometric methods to a non-physical dimension of space, the “distance” between ranks on the waiting list. This can be approximated by using the listing dates and clinical urgency categories of patients. We use a spatial autoregressive model structure and estimate our model using an asymptotically optimal instrumental variable estimator for cross-sectionally dependent data. To illustrate the relevance of this model, we analyse the impact of the introduction of a hypoplastic left heart syndrome surgery program on waiting times for other patients requiring intensive care facilities at the Sydney Children’s Hospital Network.

This model is very original in its application of spatial econometrics to patient rankings on a wait list as a non-physical dimension of space. There is currently no such modelling strategy in the literature. Hence, this paper also illustrates the ample possibilities for other applications of this method. The model has clear advantages in terms of its theoretical consistency and explanatory power and provides additional, clinically relevant information. Hence, this research makes both a methodological and an applied contribution to the literature.

3.2 Contributed - Medical Statistics: Applications of Statistics

Tuesday 6 September 2016 – 2pm-3pm

The efficiency of using embedded trials to investigate participant retention: an example using social pressure to affect retention in a health cohort study.

Sarah Rhodes, Sarah Cotterill
University of Manchester

Background The recruitment and retention of participants to health research presents a challenge, and there is very limited evidence to indicate which methods are effective. A suitable way to investigate recruitment and retention is a randomised controlled trial, embedded within a host research study. Until recently such studies have been rare, and despite a growing interest in recruitment trials,[1, 2] less attention has been paid to retention.[3]

Methods We undertook a retention RCT embedded within CLASSIC, a health cohort study of older people in Salford. Usual RCT standards for design and reporting were followed. 4447 participants were randomly assigned to one of two groups. All participants were sent a covering letter along with the second CLASSIC survey in a series. The intervention letter indicated that their response was being monitored, and would be communicated in future to the participant. The control group were sent a similar letter without the additional text. The primary outcome was questionnaire response.

Advantages

- Minimal cost
- No additional recruitment or consent required (Approval from LREC for amendment)
- Opportunity for rapid investigation of novel retention interventions (7mths from idea to results).
- No additional data collection
- Minimal ascertainment bias (participants unaware of the study)

Challenges

- Constrained to maximum sample size of host study
- May have little control over timing and methods of data collection and management
- The design of the host study may introduce analysis issues (multiple participants per household).

Recommendation Embedded trials offer an efficient and low cost way to test methods of retaining participants in health research, and there is great potential for their future use.

References

1. *Rick J.* Trials*.2014;15:407*
2. *Madurasinghe V.W.* Trials*.2016;17:27.*
3. *Brueton V.C.* The Cochrane database of systematic reviews*.2013;12:Mr000032.*

3.3 Contributed - Social Statistics: Outliers and clustered outcomes

Tuesday 6 September 2016 – 2pm-3pm

Detecting outliers in weighted univariate survey data

Anna Pauliina Sandqvist
KOF Swiss Economic Institute / ETH Zürich

Outliers and influential observations are a frequent concern in all kind of statistics, data analysis and survey data. Especially, if data are asymmetrically distributed or heavy-tailed, outlier detection is not clear-cut. Even more, for periodic data and data with multiple subsets, in which the distributional characteristics of each data sample may differ, the outlier detection is challenging as the technique may need to be adjusted each time. In this paper we examine various non-parametric outlier detection approaches for (size-)weighted growth rates from surveys and propose new respectively modified methods which can account better for non-normal data and particularly for altering levels of dispersion and asymmetry. As outlier detection (and treatment) involves in practice a lot of subjectivity, we pursue an approach in which as few as possible parameters need to be defined. We conduct a simulation study to compare these methods under different data specifications. Furthermore, an empirical illustration is presented.

3.3 Contributed - Social Statistics: Outliers and clustered outcomes

Tuesday 6 September 2016 – 2pm-3pm

Understanding ‘Don't Know’ Responses in Surveys Using Interviewer Paradata – A Cross National Analysis

Kingsley Purdam, David Bayliss, Joe Sakshaug
Manchester University

Analysing interviewer paradata from the European Social Survey (ESS) we examine respondent understanding of survey questions. A ‘Don't Know’ response to a survey question can be a result of: (i) ambivalence in terms of (a) being indifferent and not having an opinion either way; (b) having a conflicted attitude; (ii) being uncertain due to not having the knowledge to provide an opinion; and (iii) having a directional view but not wanting to express it. Research has examined the nature of ‘Don't Know’ responses in surveys using branching and follow up questions (Malhotra et al. 2009; Tourangeau et al. 2000; Sturgis et al. 2012). The removing of ‘Don't Know’ options from attitude scales has been shown to encourage some respondents to provide directional responses that they would have otherwise withheld. Contingent valuation methods have identified how the selection of ‘non opinion’ options can vary by level of educational qualification (Krosnick et al. 2002).

We examine the patterns of ‘Don't Know’ responses in the ESS in relation to political attitudes measured on uni-dimensional scales and by key demographics and context. Observation paradata from interviewers (including how well they thought respondents understood the questions, whether they were reluctant to answer and whether the respondent received any help) is then modeled to identify the meaning of a ‘Don't Know’ response. Our findings suggest that respondents answer ‘Don't Know’ across factual, value and attitudinal questions. Substantial numbers of respondents can be reluctant to answer and seek clarification before answering. A ‘Don’t Know’ response to attitudinal questions is a valid response for many, particularly in the context of evidence of increased ideological division and political party re-alignment. Paradata on how respondents answer questions can add to the accuracy of survey data by providing an analytical basis for understanding the different reasons why people answer ‘Don't Know’.

3.3 Contributed - Social Statistics: Outliers and clustered outcomes

Tuesday 6 September 2016 – 2pm-3pm

A generalization of Chao's lower bound estimator for zero-truncated one-inflated count data with an application to domestic violence data

Dankmar Boehning
Southampton Statistical Sciences Research Institute

For zero-truncated count data, as they typically arise in capture-recapture modelling, the nonparametric lower bound estimator of Chao is a frequently used estimator of population size. It is a simple, nonparametric estimator involving only counts of one and counts of two. The estimator is asymptotically unbiased if the count distribution is a member of the power series family and is providing a lower bound estimator if the distribution is a mixture of a member of the power series family. However, if there is one-inflation Chao's estimator can severely overestimate as we show here. This illustrated by routinely collected country-wide data on family violence in the Netherlands. A new lower bound estimator is developed which involves only counts of two and three, thus avoiding the overestimation caused by one-inflation. We show that the new estimator is asymptotically unbiased for a power series distribution with and without one-inflation and provides a lower bound estimator under a mixture of power series distributions with and without one-inflation. For all estimators adjusted versions are developed that reduce the bias considerably when the sample size is small.

3.4 Contributed - Data Science: Applications to online data

Tuesday 6 September 2016 – 2pm-3pm

PageRank and the Bradley–Terry model

David Selby, David Firth
University of Warwick

Network centrality measures are used to compare nodes according to their importance. Applications include ranking sports teams, ordering web pages in search results and estimating the influence of academic journals.

Eigenvector-based metrics such PageRank derive these measures from the stationary distribution of an ergodic Markov chain, whereas techniques such as the Bradley–Terry model treat ranking as a statistical estimation problem.

By using a quasi-symmetry representation, we show that the PageRank vector, suitably scaled, is a consistent estimator for the Bradley–Terry model. Scaled PageRanks can therefore be used, for example, to initialise iterative algorithms for Bradley–Terry maximum likelihood estimation, improving their performance on large datasets. We study the variance of scaled PageRank as an estimator, and find full asymptotic efficiency in some balanced situations of practical importance.

3.4 Contributed - Data Science: Applications to online data

Tuesday 6 September 2016 – 2pm-3pm

Using machine learning techniques to clean web scraped price data via cluster analysis

Matthew Mayhew, Joseph Winton
Office for National Statistics

ONS are investigating the use of prices obtained by automated collection from supermarket websites ('scraping') to compile price indices. These data are mapped into the COICOP classification structure by a machine learning algorithm based on the product name in the raw data. This can lead to misclassification within the products. Our talk describes an approach to reduce the number of misclassifications by utilizing a clustering algorithm designed to identify them. We present preliminary results on the impact to the distributions within each of the 35 grocery categories for which this price information was scraped.

3.4 Contributed - Data Science: Applications to online data

Tuesday 6 September 2016 – 2pm-3pm

Online syndromic surveillance of gastroenteritis

Elizabeth Buckingham-Jeffery
University of Warwick

Flusurvey is a web-based community cohort survey based in the UK to monitor influenza-like illness. The participants of Flusurvey are asked to self-report symptoms. Each symptom report asks for details of symptoms experienced in the past week, including diarrhoea and vomiting. These are the principal symptoms of gastroenteritis; one of the most common illnesses worldwide. However, the burden of gastroenteritis on the population is not well understood. Here we report on the potential for syndromic surveillance of gastroenteritis using the data from Flusurvey reports.

We used symptom reports for diarrhoea and vomiting submitted to Flusurvey to construct a rate of self-reported syndromic gastroenteritis. We compared this to existing measures of gastroenteritis, and gastroenteritis-causing pathogens. We computed the cross-correlation, with a confidence interval from a parametric bootstrap, of the different measures of gastroenteritis for comparison. We additionally used the Flusurvey data to give an estimate of the proportion of community cases of gastroenteritis that seek healthcare advice.

The surveillance systems analysed showed different trends in gastroenteritis rates and different age stratifications. There were only low correlations between the different measures of gastroenteritis. 15% of Flusurvey respondents with gastroenteritis symptoms sought medical attention.

This work shows that Flusurvey, the internet-based surveillance system for influenza-like illness, can also be used to give some measure of gastroenteritis incidence and an estimate of the usage of healthcare services by those with gastroenteritis symptoms. The Flusurvey system provides additional information on gastroenteritis burden, in particular of cases where patients do not seek healthcare services, but further surveillance is required if the burden of gastroenteritis in the community is to be fully understood.

This work was completed with Thomas House from the University of Manchester, the Real-time Syndromic Surveillance Team at Public Health England, and the Flusurvey team at the London School of Hygiene and Tropical Medicine.

3.6 Contributed - Official Statistics: Inflation

Tuesday 6 September 2016 – 2pm-3pm

Drivers of measured, perceived and expected inflation

Jens Mehrhoff
Deutsche Bundesbank

The official measure of a Consumer Price Index is based on a representative concept, which does not necessarily coincide with the individual consumer's subjective perception of price dynamics. This paper explains why consumer surveys often show that people 'feel' inflation to be higher than the actual price indices indicate and what forms people's perceptions of inflation.

Using micro data on household expenditures from the German Income and Consumption Survey, we compute income-specific price indices. We find differences in the measured inflation rate across income groups, which are mainly related to different expenditure shares in the food and energy component and the fact that high-income households dominate the weight pattern.

Likewise, households's formation of inflation perceptions and expectations is strongly determined by socio-demographic characteristics. Based on a monthly survey of about 2,000 German households, we find via an ordered logit model that consumers perceive a higher rate of inflation with increasing age, household size, and when the respondent is female or unemployed. In contrast, higher education and household net income are linked to a lower perceived rate of inflation. Thus, we are able to show the distribution of inflation perceptions by household net income.

Extending the study beyond Germany, we demonstrate that perceived inflation and the current rate of inflation move broadly in line. However, in 2002 a prolonged impact of the euro cash changeover on the perceptions came to light in several countries. In addition, country differences in the inflation perceptions cannot be fully explained by inflation differentials in the euro area. Hence, there is some kind of a structural 'level' effect at play in Member States.

Last, besides socio-demographic determinants and the inflation perception itself, we also show that oil price changes and the European Central Bank's (ECB) monetary policy measures are significant drivers of inflation expectations in Germany.

3.6 Contributed - Official Statistics: Inflation

Tuesday 6 September 2016 – 2pm-3pm

Modelling and Forecasting Inflation Rate Volatility

Paul Kattuman, Christoph Weiss
University of Cambridge

The adverse effects of inflation volatility on economic growth and welfare are well known. But the generating process that underlies inflation volatility is not as well understood as it should be. Using monthly data that underlies the Retail Price Index for the U.K., we analyse the drivers of the inflation rate, and its volatility, for the period since inflation targeting began.

We establish the value of hierarchical time series modelling for forecasting both the volatility of the aggregate inflation rate, as well as the inflation rate itself. We explicitly incorporate the scheme of aggregation whereby product-level inflation rates combine into category level inflation rates, and in turn into the overall inflation rate; and also distinguish between the common, category and idiosyncratic components at the product level.

In modelling aggregate inflation rate volatility, we explicitly consider the variances and the co-variances among the basic constituents that aggregate up into the inflation rate. These are the common, category and idiosyncratic components of the product level inflation rates. We find that aggregate inflation volatility closely tracks the pattern in the covariance component, and that this is mainly driven by the variances of common shocks shared by all products, and the co-variances between idiosyncratic, product-level shocks. Category level shocks play a subsidiary role.

A forecasting system for the mean and for the variance, that exploits the index structure of the aggregate inflation rate, by using the disaggregated hierarchical time series forecasting framework provides both more accurate forecasts and greater insight into the inflation process. This is evident from the accuracy comparisons of this approach with conventional univariate modelling approaches.

4.1 Contributed - Methods & Theory: Bayesian Time Series Analysis

Tuesday 6 September 2016 – 3.10pm-4.10pm

A Bayesian approach to measuring the non-stationarity of a time series

Sourav Das, Guy Nason
University of Bristol

Since the 1960's non-stationary time series have been investigated extensively. Methodology and theory have evolved rapidly since Dahlhaus' construction of locally stationary processes in the 1990's. Much of the theory in above constructions rely on assumptions of smoothness on the time varying transfer function. However when modelling real data, tools for assessing such regularity conditions are yet to be developed. We propose a methodology for measuring the degree of non-stationarity in a time series. We use principles of non-parametric regression to measure the roughness of posterior distribution of the underlying signal and show its association with stationarity. The method is tested on simulated time-varying autoregressive processes. We also demonstrate potential applications of this index in finding solutions to pertinent questions in Earth Sciences.

4.1 Contributed - Methods & Theory: Bayesian Time Series Analysis

Tuesday 6 September 2016 – 3.10pm-4.10pm

Bayesian Quantile Regression for Discrete Data

Xi Liu, Keming Yu
Brunel University London

Objectives of the work:

Quantile regression for continuous response, including Bayesian inference quantile regression, has been widely developed in literature and applied in practice. But there is little research on quantile regression for discrete data in literature, particularly from a Bayesian perspective. Discrete data are common in many disciplines. Regression analysis of discrete data has been an active and promising area of research. Discrete data are often analyzed incorrectly with ordinary least squares regression.

The methods:

This paper introduces Bayesian quantile regression for discrete data via a discrete asymmetric Laplace distribution. The method provides a direct Bayesian approach for discrete data with natural and easy interpretation of the regression results. The posterior distribution under this approach is shown not only consistent irrespective of the original distribution of the response but also proper with regarding to improper priors for the unknown model parameters.

The results

The method is shown robust and consistent numerically and theoretically. The Bayesian approach which is fairly easy to implement
provides complete univariate and joint posterior distributions of parameters of interest. The posterior distributions of the unknown model parameters are obtained by using MCMC methods implemented in R. Although we have chosen improper flat priors in our numerical analyses, there is scope of using other priors in a relatively straightforward fashion. The extensions to spatial and random effects models would be represent interesting areas of development.

.

4.1 Contributed - Methods & Theory: Bayesian Time Series Analysis

Tuesday 6 September 2016 – 3.10pm-4.10pm

Bayesian outlier detection in non-Gaussian AutoRegressive time series

Maria Eduarda Silva, Isabel Pereira
Universidade do Porto & CIDMA

Observations that look discordant from most observations in a data set are often encountered in time series. Neglecting the presence of such outliers hinders meaningful statistical inference, leading to model misspecification, biased parameter estimation, and poor forecasts. Several methodologies for detecting and estimating outliers and other intervention effects have been established in the framework of Gaussian linear time series with emphasis on iterative procedures and likelihood based statistics. Apparently there is not much work, so far, on the analysis of outliers in positive-valued and count time series which arise in a wide variety of fields. These data are naturally non Gaussian and typically right-skewed, causing a need for especially designed models and procedures. A useful class of models for positive valued and count time series are the autoregressive models with margins in the convolution closed infinitely divisible class and non negative serial correlation. In this work we consider the problem of modelling outliers in non Gaussian AR(1) processes, focussing on the integer-valued time series models. These processes although simple are flexible enough for approximating the dependence structure in many positive-valued and count time series. The context is a retrospective analysis of purely additive outliers occurring at unknown time locations and not entering the dynamics of the model. The Bayesian approach proposed here does not require beforehand knowledge on the number and location of outliers in the series and treats equally all the observations, outlying or not. Moreover, this approach provides at each time point, a probability of outlier occurrence and an estimate of the corresponding size. The methodology is illustrated with synthetic Poisson integer-valued autoregressive time series and an observed time series.

4.2 Contributed - Medical Statistics: Clinical Trials

Tuesday 6 September 2016 – 3.10pm-4.10pm

Complexities of running mediation analysis in a randomised controlled trial with two active treatment arms

Katy Sivyer, Elizabeth Allen, Rebecca Murphy, Zafra Cooper, Christopher Fairburn
University of Oxford

Mediation analysis is ideally undertaken under experimental conditions comparing an active treatment condition to an inactive control. Such designs enable the researcher to isolate the relationships between treatment, mediator and outcome to understand how treatments work. Randomised controlled trials of psychological treatments are time-consuming and resource intensive. Mediation analysis embedded within a randomised controlled outcomes trial can capitalise on research resources by combining outcomes research with research into the purported underlying mechanisms of the treatment. However, the increasing use of non-inferiority, equivalence and superiority designs in randomised controlled treatment trials means that the use of inactive control conditions may become less common. Conversely, there is an increasing focus on understanding the mechanisms of actions of treatment, particularly psychological treatments, given that little is understood about how these treatments work. This talk will discuss the complexities of running mediation analysis in a randomised controlled trial of two psychological treatments.

4.2 Contributed - Medical Statistics: Clinical Trials

Tuesday 6 September 2016 – 3.10pm-4.10pm

How early stopping for futility in clinical trials can affect the estimated treatment effect

Stephen Walter
McMaster University

We consider the estimation of the treatment effect in clinical trials whose protocol includes interim analyses, and possible early termination of the study for reasons of futility – when the early data suggest only limited treatment benefit. We calculate the interim conditional power, using alternative assumptions about the future data. We then derive expressions for various parameters of interest:

- the expected treatment effect in studies that stop early;
- the expected treatment effect in completed studies, i.e. that do not stop early;
- the overall estimation bias associated with a futility stopping rule; and
- the probability of stopping at an interim analysis.

We evaluate these expressions for typical trial scenarios, and show that these parameters depend on several factors, including the true treatment effect, the effect size that the trial was designed to detect, study power, the number of planned interim analyses, and what is assumed about future data. The overall bias is often small, but there may be substantial under-estimation of the treatment effect in studies that actually stop early.

We illustrate these ideas using data from two trials (concerning treatments for lung and breast cancer) that stopped early for futility. The first trial may have had substantial conditional power at the interim analysis, and its interim effect size may have been a substantial under-estimate. The second had low conditional power, but it still experienced under-estimation of the treatment effect.

Investigators should be aware of potential mis-estimation of the treatment effect when considering stopping for futility, particularly in trials with multiple interim analyses, or with small sample sizes and numbers of events.

4.2 Contributed - Medical Statistics: Clinical Trials

Tuesday 6 September 2016 – 3.10pm-4.10pm

Outcome selection in clinical trials – looking back at the problems and moving forward with solutions

Paula Williamson, Jamie Kirkham, Carrol Gamble, Kerry Dwan
University of Liverpool

Selection of outcomes to measure in trials designed to compare different interventions is crucial. It has been estimated that less than half of all outcome data collected in trials is fully reported, with missing data due to unpublished trials, poor reporting, and choosing not to include particular results within trial reports.

Difficulties caused by heterogeneity in outcome measurement across studies are well known. Empirical research provides strong evidence that outcome reporting bias (ORB), defined as the results-based selection for publication of a subset of the recorded outcome variables, is a significant problem in a quarter to a third of randomised trials and can have major impact in a fifth of systematic reviews. In interviews, trialists seemed unaware of the implications for the evidence base of not reporting all outcomes and protocol changes.

Systematic reviewers facing these challenges should contact trialists to try to obtain the missing data. They may subsequently apply a statistical approach as part of a sensitivity analysis. Bias bound estimation, multivariate meta-analysis, and modelling the selection process have been proposed.

Trial registration and improved reporting should help to reduce ORB, but for findings to influence policy and practice, outcomes chosen for measurement need to be relevant to patients, public, healthcare professionals and others making decisions about health care. So much could be gained if an agreed core outcome set (COS) of appropriate and important outcomes was measured and reported in all clinical trials of effectiveness in a specific condition. The COMET (Core Outcome Measures in Effectiveness Trials) Initiative, http://www.comet-initiative.org/, an innovative global project, brings together people interested in COS development and application.

This talk will review progress made with both statistical and non-statistical solutions to this problem.

4.3 Contributed - Social Statistics: Polls and Europe

Tuesday 6 September 2016 – 3.10pm-4.10pm

Are Regions More Important than Countries in Analyses of Eurostat Surveys?

Neil Spencer, Sahib Matharu
University of Hertfordshire

Data on businesses, employment and social affairs in the European Union are collected by national governments under the co-ordinating eye of Eurostat who provide a unified methodology for data collection.

As well as collating the data and providing summary statistics that can be used for international comparisons, Eurostat enable a number of the surveys to have their data released to authorised researchers in anonymised form. These microdata can then be analysed to deepen understanding of the interplay between aspects of the survey, giving the potential to influence policy at national and European levels. Analyses of these microdata frequently use multilevel modelling techniques and it is natural and convenient to use countries as a level in the model.

However, for multilevel modelling to be successful, it is important that an appropriate grouping structure is used and despite the convenience of using countries, it is not clear that this is the most appropriate clustering to use. Convergence of business and employment law across the European Union means that although differences still exist, they are not as great as they might otherwise be. Additionally, if country is the only grouping structure being used in the modelling, the grouping of individuals within regions of a country is being ignored. The omitting of levels in a multilevel model is known to potentially affect the results obtained and conclusions drawn.

This work investigates the potential influence that ignoring regions may have on analyses of Eurostat microdata by considering the intra-class correlation coefficient (ICC). Using publicly available data, a range of ICCs are calculated for different potential levels of the multilevel model. The possible impact they may have on analyses if ignored are assessed.

4.3 Contributed - Social Statistics: Polls and Europe

Tuesday 6 September 2016 – 3.10pm-4.10pm

Predicting the 2016 UK EU membership referendum

Timothy Martyn Hill
LV

- **Objective**: Judge the predictive ability of predictions for the 2016 UK EU membership referendum
- **Method**: Collate predictors (polls, models, betting odds and spreads), discuss and establish a method of common measurement, track them over time
- **Results**: List how successful each method was and if/when they became useful as predictors
- **Other**: This is the second of a series published by the author in Significance. The first set was published in 2015 (and presented at Conference) and covered the UK General Election, the third covers the 2016 US Presidential and is scheduled for 2017. Interest has also spread to the Elections, Public Opinion and Parties (EPOP) specialist group of the Political Studies Association

4.5 Contributed - Spatial Statistics: Modelling of environments

Tuesday 6 September 2016 – 3.10pm-4.10pm

Estimating changes in earthquake occurrence rates

Rakesh Paleja, Stijn Bierman, Tim Park
Shell Global Solutions

Gas production from the Groningen field located in the north of the Netherlands induces earthquakes that are causing a concern with the local population. In order to address the problem, the production of gas from the field has been reduced since 2014. Typically, the number of earthquakes that occur in a given time is Poisson distributed. Thus, detecting changes in earthquake rates is often difficult and a challenging task when the count of earthquakes is small.

In this conference presentation we will discuss the use of spatio-temporal statistical techniques to understand the behaviour of earthquakes in the Groningen field. We find that the earthquake rate prior to 2014 varies in both time and space for the entire Groningen field. Therefore, in order to detect if the earthquake rate after 2014 has reduced, we divided the earthquake catalogue into epochs such that the earthquake rate in each epoch is stationary. The inter-event times (IITs) for each epoch is then computed using a Bayesian approach. Thus, the posterior density of the IITs post and pre 2014 can be compared. This novel method shows that since 2014, there was a statistically significant reduction in the earthquake occurrence rate for the Loppersum region of the Groningen field. The results are consistent with conventional, rate based, statistical tests used by seismologist. The Bayesian method we propose offers an advantage in that if the earthquake count data is from an over-dispersed Poisson distribution, the likelihood function can be adapted and the posterior IIT can be reliably estimated.

4.5 Contributed - Spatial Statistics: Modelling of environments

Tuesday 6 September 2016 – 3.10pm-4.10pm

Does increasing road lighting increase road injuries?

Paul Marchant
Leeds Beckett University

The work presented examines the effect of increasing road lighting on reported personal injury road traffic accidents, in one large city. The increase in lighting occurs by brightening and whitening road lamps. Other work suggests that new Private Finance Initiative (PFI) lighting has not reduced crime by the 20% that was claimed it would. This presentation examines the impact on road accidents when tens of thousands of lamps have been changed and tens of thousands of road injuries have occurred during the 9-year time period of the study.

A multilevel analysis of the time series of the road traffic accident data will be presented. The analysis, done at the small area level, utilises the times of lighting change and the times of occurrence of the accidents. The analysis recognises that the underlying accident rate is simultaneously changing (generally decreasing) over time, in the absence of any change in the lighting level. The analysis follows its protocol with some small variations, which will be discussed.

The results are interesting in that they suggest a modest but detectable increase in the accident rate as the lighting increases. This effect remains despite considerable checking and discussion of the methods and analysis!

4.5 Contributed - Spatial Statistics: Modelling of environments

Tuesday 6 September 2016 – 3.10pm-4.10pm

Modelling blinking fluorophores in Super Resolution Microscopy

Lekha Patel, Edward Cohen
Imperial College London

Super-resolution microscopy is a collection of imaging techniques allowing experimenters to delve beyond classical resolution limits to image cellular structures in the nanometre scale. The key element to the success of super-resolution techniques is the stochastic blinking of fluorophores (light emitting molecules) allowing sparse subsets to be localised with very high precision. These localisations collected across time build a spatial point pattern of molecular positions. However, with multiple blinks for each fluorophore, these point patterns can be a misleading representation of the spatial organisation of molecules and thus understanding this blinking process is crucial to proper inference. We start by modelling an individual fluorophore as a continuous-time homogeneous Markov process with an absorbing or permanently dark state corresponding to photo-bleaching characterised by unknown transition rates. In addition, we define the imaging process allowing one to observe localisations of this fluorophore in a given photographic frame. We thereby formulate a Hidden Markov Model (HMM) to link the discrete time observed process with its continuous time signal and thus develop an estimation procedure for the transition rates between states. This consequently provides us with a basis to investigate the distributive properties of localisations from a molecule across a series of frames. We then discuss a multivariate extension of this method to allow for the more common scenario of multiple emitters. Finally, we perform a simulation study to analyse the effectiveness of our model.

4.6 Contributed - Official Statistics: Measuring school performance and statistical literacy

Tuesday 6 September 2016 – 3.10pm-4.10pm

An Indicator for Statistical Literacy based on National Newspaper Archives

Thilo Klein, Anais Galdin, El Iza Mohamedou
OECD/PARIS21

This paper develops and reports on a composite indicator for statistical literacy as part of the Busan Action Plan for Statistics (BAPS) logical framework. The BAPS was endorsed at the Fourth High-Level Forum on Aid Effectiveness in Busan, Korea in 2011 with the objective to "fully integrate statistics into decision-making". Statistical literacy is a prerequisite to effectively use statistics to inform decisions for planning, analysis, monitoring, and evaluation, thus increasing transparency and accountability. The indicator was developed in collaboration with a task team set up by the Partnership in Statistics for Development (PARIS21).

As a theoretical framework, we adapted the taxonomy of the statistical literacy construct of Watson and Callingham (2003). To measure statistical literacy empirically, we turn to keywords related to statistics and statistical fallacies in national newspaper articles based on this taxonomy. This is essentially for three reasons. First, the writing of journalists can be seen as a reflection for a nation's demand for statistical facts and depth of critical analysis. Second, newspaper articles are generally available, which makes them representative of a country's literate population and easily accessible for text analysis. Third, regional numeracy assessments are reported infrequently and often not comparable across countries.

The purpose of the indicator is to set and monitor targets and report on them annually for developing and developed countries. To date, we have analysed a total of 56,231 articles in general news (taxonomy level 1) and 5,589 articles that report on studies and research findings (taxonomy levels 2 and 3). For each of the levels of statistical literacy, the resulting score is given as the percentage of articles that contain at least one search term from the compiled keyword lists. The results are presented by language groups (English, French, Spanish, Portuguese) to allow for a direct comparison between countries.

4.6 Contributed - Official Statistics: Measuring school performance and statistical literacy

Tuesday 6 September 2016 – 3.10pm-4.10pm

Evaluating the Provision of School Performance Statistics: A Two-Sided Matching Model of School Choice

Thilo Klein
OECD/PARIS21

The value of statistics is evident but hard to measure, which makes investments in data often difficult to justify. This paper measures the value of school performance information in school choice. This context is particularly suitable because (i) it is an important decision that sets the course for a child's social and economic development and (ii) it allows economists to infer parents' use of publicly available information as well as its impact on schooling outcomes.

The inference in this paper is based on detailed administrative data on preferences and outcomes of more than 110,000 students and 2,000 schools in Hungary. To analyse this data, I develop a novel two-sided matching model that allows for a more structural approach than extant literature by jointly estimating both student and school preferences together with schooling outcomes.

The paper makes three methodological contributions. First, the model allows for the estimation of both student and school-side preferences. Previous research has not accounted for school-side preferences. This clearly overstates the value of statistics for parents, i.e. it is not helpful to know what is a good school if their children cannot get in. Second, the model does not require truth-telling. Non-truth-telling is a major concern because it is rational for students not to rank top schools if they anticipate that they will not be admitted anyway. If these preference rankings are taken at face value, low-performing students will be falsely seen as demanding low-performing schools. Third, causality in the schooling outcome equation is established using a novel instrumental variable. The exclusion restriction is that the characteristics of all agents in the market affect which student attends which school, but a student's performance is independent of other schools and their students.

4.7 Contributed - Industry & Commerce: Compression, Compaction and Visualisation

Tuesday 6 September 2016 – 3.10pm-4.10pm

Compression of High Frequency Acoustic Data using Wavelet Analysis

Shiraz Basheer, Tim Park, Rakesh Paleja
Shell

An oil producing well can stretch as deep as 12km underground. To get an idea of the flow composition in the pipe we currently have to wait for the flow to reach the surface. Distributed Acoustic Sensing (DAS) however uses optical fibres to provide near real-time measurements of acoustic and thermal disturbances along the length of a well. Using this technique any change in flow composition can be estimated in real time. Data is collected at a very high frequency of 10,000 Hz and at regularly spaced depths. The rate of data acquisition is 1GB per min which means after 15 minutes it would fill an iPhone. Storage of such a huge amount of data is challenging, especially in field locations, and so compression becomes important. We considered different lossy compression methods such as Robust PCA, SVD and Wavelet Compression. These mainly focused on removing the noise component and retaining only significant information, however it is also important that we are still able to perform inference on the compressed data. This inference typically consists of spectral methods such as Fourier analysis alongside multivariate techniques such as PLS regression to predict the flow rate and composition throughout the well. We also consider the possibility of compressing the data as it is acquired. In this talk we present different compression methods applied to real well data and compare their impact on flow rate predictions.

4.7 Contributed - Industry & Commerce: Compression, Compaction and Visualisation

Tuesday 6 September 2016 – 3.10pm-4.10pm

Statistics with a Human Face- Visualisation

Liberty Vittert, Adrian Bowman, Stanislav Katina
University of Glasgow

Three-dimensional surface imaging, through laser-scanning or stereo-photogrammetry, provides high-resolution data defining the surface shape of objects. Using a human face as this object, each image corresponds to an observation, a manifold, represented by a triangulated point cloud. In an anatomical setting this can provide invaluable quantitative information. Particular applications vary widely including success or failure of cosmetic/reconstructive plastic surgery, facial recognition, facial asymmetry, concepts of sexual dimorphism, and even the survival of mussels (food we consume) given climate change. However, the initial challenge is to characterize these complex surfaces, without laborious manual intervention. Surface curvature provides the key information in doing this, allowing for a creating of a surface “mask” replicable throughout all these objects. Once the full surface representation has been obtained, the new issue arises of how to best characterize and visualize the differences in shape. The issues involved with analysis of this data and multiple visualization methods will be discussed and illustrated.

4.7 Contributed - Industry & Commerce: Compression, Compaction and Visualisation

Tuesday 6 September 2016 – 3.10pm-4.10pm

Estimation of reservoir compaction using surface displacement measurements

Stijn Bierman, Saptarshi Das
Shell Global Solutions NL

The extraction of pressurised fluids and gas from an oil or gas reservoir may cause the rock in the reservoir to compact due to pore volume reduction. Compaction of the rock results in displacements (for example subsidence) at the surface which may be measured using, for example, optical levelling techniques, Global Position System (GPS) or Interferometric Synthetic Aperture Radar (InSAR) data. We present statistical methodology that may be used to estimate compaction in an oil or gas reservoir using such surface displacement measurements. The independent contribution of compaction in different parts of the reservoir to the observed surface displacements cannot be determined without imposing some form of regularisation. Regularisation is achieved by imposing spatial smoothness on the estimates of compaction in the reservoir. A particular challenge is to account for the potential presence of spatio-temporally correlation in the model residuals. An overview of the geomechanical aspects of the problem as well as the statistical modelling approached will be presented.

5.1 INVITED - Methods & Theory: Intractable Likelihood

Tuesday 6 September 2016 – 4.30pm-5.50pm

Bayesian nonparametric approaches to quantifying dependence between random variables

Sarah Filippi, Chris Holmes, Luis Nieto Barajas
University of Oxford

Nonparametric and nonlinear measures of statistical dependence between pairs of random variables have proved themselves important tools in modern data analysis, where the emergence of large data sets can support the relaxation of linearity assumptions implicit in traditional association scores such as correlation. In this talk, I will present two Bayesian nonparametric procedures to test for dependences. In the first one a tractable, explicit and analytic quantification of the relative evidence of dependence vrs independence, using Polya tree priors on the space of probability measures which can then be embedded within a decision theoretic test for dependence. In the second procedure the unknown sampling distribution is specified via Dirichlet Process Mixtures (DPM) of Gaussians, which provide great flexibility while also encompassing smoothness assumptions. After describing the methods, I will contrast their performances in high dimensional spaces. These procedures can accommodate known uncertainty in the form of the underlying sampling distribution and provides an explicit posterior probability measure of both dependence and independence. Well known advantages of having an explicit probability measure include the easy comparison of evidence across different studies, the inclusion of prior information, and the integration of results within formal decision analysis.

5.1 INVITED - Methods & Theory: Intractable Likelihood

Tuesday 6 September 2016 – 4.30pm-5.50pm

Scaling MCMC algorithms to big data problems using parallel computing

Chris Nemeth, Chris Sherlock
Lancaster University

MCMC has become one of the most popular algorithms for analysing complex Bayesian models. Unfortunately, standard implementations of MCMC do not scale well to the 'big data' scenario, where millions of observations need to be evaluated at each iteration making the algorithm prohibitively slow.

As computational power becomes cheaper, and the availability of parallel processors increases, it's natural utilise this cheap computational power to perform MCMC in parallel. We can split the data across multiple machines and running MCMC in parallel. However the challenge is to then recombine the posterior distributions to form the full posterior.

In this talk I'll give a review of some techniques that have already been applied and talk about some recent we have done to address this problem using Gaussian Process models.

This is joint work with Chris Sherlock.

5.1 INVITED - Methods & Theory: Intractable Likelihood

Tuesday 6 September 2016 – 4.30pm-5.50pm

Small ScaLE and Large ScaLE: Developments in the Scalable Langevin Exact Algorithm

Murray Pollock, Gareth Roberts, Paul Fearnhead, Adam Johansen
University of Warwick

We present a new methodology for exploring without error posterior distributions by modifying methodology for exactly simulating diffusion sample paths (the Scalable Langevin Algorithm (ScaLE)), in this talk we will present applications of this new methodology beyond “Big Data” to other settings, such as large dimensional models.

5.2 INVITED - Medical Statistics: Evaluation of non-randomised health policy interventions

Tuesday 6 September 2016 – 4.30pm-5.50pm

Assessing the impact of Be Clear on Cancer campaigns – which methods are best?

Carolynn Gildea
Public Health England

Background

Public Health England, in partnership with the Department of Health and NHS England have run a number of 'Be Clear on Cancer' (BCOC) campaigns since early 2011. The campaigns aim to highlight the signs and symptoms of a range of cancers, to encourage people with the relevant signs and symptoms to visit their GP and so help improve early cancer diagnosis rates.

Objective

An important aspect of running such campaigns is to evaluate the impact and effectiveness of them. However, evaluation is complicated by a number of factors – including aspects of how the campaigns were structured, that evaluation relied largely on routine and administrative data, and the wider cancer and health context.

Methods

The talk will focus on one of a number metrics considered within the evaluation of the Be Clear on Cancer campaigns – suspected cancer referrals to secondary care. Several approaches have been considered to evaluate the campaigns' impact, including examination of trends; consideration of changes over time with 'differences in difference'; linear models with prediction; and interrupted time series.

Discussion

The talk will explore some of the benefits and limitations of the methods considered for this work, and therefore highlight some of the issues faced when evaluating non-randomised interventions.

5.2 INVITED - Medical Statistics: Evaluation of non-randomised health policy interventions

Tuesday 6 September 2016 – 4.30pm-5.50pm

Propensity score matching for selection of local areas as controls for evaluation of effects of alcohol policies in case series and quasi case-control designs.

Frank de Vocht, Rona Campbell, Alan Brennan, John Mooney, Colin Angus, Matthew Hickman
University of Bristol

Objectives: In the context of public health policy making and implementation it is generally not possible to conduct randomized controlled trials to evaluate the effectiveness of new policies, because implementation of area-level policies are generally not under the control of the researchers. One method often used in non-randomized epidemiological studies of individuals is the use of propensity score matching (PSM) to match intervention to the most appropriate controls to approximate the counterfactual. We aimed to assess PSM at aggregated local authority area (LAU) to match cases with a alcohol policy intervention to control LAUs with respect to alcohol-related baseline covariates, so that the evaluation of the public health impact of the an alcohol policies will be minimally biased. Two different study design are described for which PSM can be useful: (1) prospective evaluation of alcohol policies, and (2) a novel two-stage quasi case-control design.

Methods: Alcohol-related indicator data (Local Alcohol Profiles for England, PHE Health Profiles and ONS data) were linked at LAU level. Six LAUs (Blackpool, Bradford, Bristol, Ipswich, Islington, and Newcastle-upon-Tyne) as sample intervention or case areas were matched to two control LAUs each using PSM. For the quasi case-control study a second stage was added aimed at obtaining maximum contrast in outcomes based on propensity scores.

Conclusions: Using PSM all LAUs were successfully (i.e. satisfied all matching criteria) to control areas such that covariate distributions were comparable at baseline. To construct the novel quasi case-control study, in a second stage the stage 1 PSM LAUS sets of LAUs were again matched based on maximum variability in health outcomes. The use of PSM for area-level alcohol policy evaluation, but also for other public health interventions, will improve the value of these evaluations by objective and quantitative selection of the most appropriate control areas in non-randomized health policy interventions.

5.2 INVITED - Medical Statistics: Evaluation of non-randomised health policy interventions

Tuesday 6 September 2016 – 4.30pm-5.50pm

Regression based quasi-experiment when randomisation is not an option: interrupted time series analysis

Evangelos Kontopantelis
University of Manchester

Randomised controlled trials (RCTs) are considered the ideal approach for assessing the effectiveness of interventions. However, not all interventions can be assessed with an RCT, whereas for many interventions trials can be prohibitively expensive. In addition, even well designed RCTs can be susceptible to systematic errors leading to biased estimates, particularly when generalising results to “real world” settings.

Observational studies can address some of these shortcomings, but the lack of researcher control over confounding variables and the difficulty in establishing causation mean that conclusions from studies using observational approaches are generally considered to be weaker. However, with quasi-experimental study designs researchers are able to estimate causal effects using observational approaches.

Interrupted time series (ITS) analysis is a useful quasi-experimental design with which to evaluate the longitudinal effects of interventions, through regression modelling. The term quasi-experimental refers to an absence of randomisation, and ITS analysis is principally a tool for analysing observational data where full randomisation, or a case-control design, is not affordable or possible. Its main advantage over alternative approaches is that it can make full use of the longitudinal nature of the data and account for pre-intervention trends. This design is particularly useful when “natural experiments” in real word settings occur, for example, when a health policy change comes into effect.

However, implementing an interrupted time-series analysis can be methodologically challenging. We will present a number of possible modelling approaches using a series of published examples of increasing complexity, and discuss their advantages, disadvantages and underlying assumptions.

5.3 Invited - Social Statistics - Confronting risk: Statistical perspectives on life, death and disaster

Tuesday 6 September 2016 – 4.30pm-5.50pm

Predictions, disasters and "Dirty Harry"

Robert Matthews
Aston University

Predicting the timing, location and intensity of disasters is an age-old dream. Yet despite huge time, effort and funding being poured into the challenge, the level of success has been mixed. While warnings of some natural disasters – notably hurricanes – are now practicable, other calamities continue to defy adequate prediction. I shall outline the nature of the challenge and what is required of a practicable prediction method. I shall then discuss why some phenomena will forever defy our efforts – and what we should do about it.

5.3 Invited - Social Statistics - Confronting risk: Statistical perspectives on life, death and disaster

Tuesday 6 September 2016 – 4.30pm-5.50pm

Statistics of disaster near-misses

Gordon Woo
RMS

Major disasters are thankfully rare, and the statistics of such destructive events in any region are sparse. No amount of resampling of historical event data can avoid the surprise of the unknown. Furthermore, extreme value analysis is sensitive to the severity of the biggest observed losses, which are prone to substantial volatility.

There is a human anthropocentric bias to regard the past as fixed, and somehow special, rather than just one of an infinitude of alternative possible realizations. The catalogue of extreme events is the outcome of underlying dynamics which are typically complex and nonlinear, and have cascading behaviour. Statistical algorithms that best explain a small number of observations in a brief time window may overfit the historical data, because what actually happened may not necessarily have been particularly likely. This gives rise to a systemic risk, whereby all risk models are tightly calibrated against a sparse set of historical observations that may be missing crucial events.

A remedy is a counterfactual approach involving stochastic modelling of the past, which is recognized as far from inevitable but haphazard. Most so-called Black Swan events have either nearly happened before, or might have happened before. But, because of outcome bias, as noted by Kahneman, inadequate attention is given to crises where disaster was narrowly averted. However, by exploring further if things had gone wrong, or had turned for the worse, insight can be gained into extreme events that can warn of future disaster modes not yet witnessed. Examples are given from a wide range of natural and man-made perils, including earthquakes, volcanic eruptions, windstorms, solar storms, industrial accidents and terrorism.

5.3 Invited - Social Statistics - Confronting risk: Statistical perspectives on life, death and disaster

Tuesday 6 September 2016 – 4.30pm-5.50pm

How risky is our intuition for risk?

Xiao-Li Meng
Harvard University

Whether we realize it or not, each of us routinely engages in intuitively assessing risks for our plans and actions, from crossing streets to “crossing the line.” A key part of this assessment is our perception of the relevant “denominator” – how many trials of my planned action would result in an outcome that I don't want to happen? Unfortunately many of our perceptions were formed from anecdotes, myths, media selective reporting, etc., making risk assessment one of the riskiest behaviors in our daily lives. Bayes theorem can help us to greatly reduce such risk, but it is inaccessible to most because of its (perceived) mathematical complexity. However, a recent “puzzler” of CarTalk, a popular radio show in Boston, demonstrated how the general public can more readily use and appreciate a “quick-and-dirty” approximation to Bayes theorem to assess probabilities of rare events. It also reinforces the critical importance of specifying the right “dominator,” that is, the reference population that is most relevant for *my* action.

5.4 INVITED - Data Science: Theory and Methods of Data Science

Tuesday 6 September 2016 – 4.30pm-5.50pm

Scalable inference for a full multivariate stochastic volatility model

Petros Dellaportas, Anastasios Plataniotis, Michalis Titsias
University College London

We introduce a multivariate stochastic volatility model for asset returns that imposes no restrictions to the structure of the volatility matrix and treats all its elements as functions of latent stochastic processes. When the number of assets is prohibitively large, we propose a factor multivariate stochastic volatility model in which the variances and correlations of the factors evolve stochastically over time. Inference is achieved via a carefully designed feasible and scalable Markov chain Monte Carlo algorithm that combines two computationally important ingredients: it utilizes invariant to the prior Metropolis proposal densities for simultaneously updating all latent paths and has quadratic, rather than cubic, computational complexity when evaluating the multivariate normal densities required. We apply our modelling and computational methodology to 571 stock daily returns of Euro STOXX index for data over a period of 10 years.

5.4 INVITED - Data Science: Theory and Methods of Data Science

Tuesday 6 September 2016 – 4.30pm-5.50pm

Bernstein-von Mises theorem and high dimensional nonidentifiable possibly misspecified models

Natalia Bochkina, Peter Green
University of Edinburgh

We consider a broad class of statistical models that can be misspecified and ill-posed, from a Bayesian perspective. This provides a flexible and interpretable framework for their analysis, but it is important to understand the relationship between the chosen Bayesian model and the resulting solution, especially in the ill-posed case where in the absence of prior information the solution is not unique.

Compared to earlier work about the Bernstein-von Mises theorem for nonregular well-posed Bayesian models, we show that non-identifiable part of the likelihood, together with the constraints on the parameter space, introduce a more complex geometric structure of the posterior distribution around the best reconstruction point in the limit, and provide a local approximation of the posterior distribution in this neighbourhood.

The results apply to misspecified models which allows, for instance, to evaluate the effect of of model approximation on statistical inference.

Emission tomography is taken as a canonical example for study, but our results hold for a wider class of generalised linear inverse
problems with constraints.

5.4 INVITED - Data Science: Theory and Methods of Data Science

Tuesday 6 September 2016 – 4.30pm-5.50pm

Topological Data Analysis at the ATI

Ulrike Tillmann
University of Oxford

The talk will introduce topological data analysis (TDA). The main goal is to give the audience a sense of this relatively new area of research, its challenges and its opportunities. Drawing on classical geometry and topology, TDA provides a flexible tool generating global statistics for data that has been applied to a large array of domains, including medicine, computer vision, and manufacturing. Time permitting we will touch on some of the research questions that will be persued at the Alan Turing Institute.

5.5 INVITED - Spatial Statistics: Spatial Statistics for tropical disease control

Tuesday 6 September 2016 – 4.30pm-5.50pm

Mapping progress in malaria control with spatial statistics (with application to a large scale serological survey from Cambodia)

Ewan Cameron
University of Oxford

The past 15 years have seen a substantial decline in the prevalence and clinical incidence of malaria illness across the endemic regions of sub-Saharan Africa, Asia and South America owing primarily to massive investment in the key interventions (ITNs, ACTs, and IRS). As transmision declines it is essential for national malaria control programs to supplement their routine parasite prevalence surveys with more sensitive diagnostics of exposure that can accurately quantify local transmission intensity in the pre-elimination regime. For this reason there is much interest in the potential for serological markers, which trace historical malaria exposure in the host immune response, to serve this purpose. In this talk I will describe progress towards incorporating serological data into statistical map making by way of discrete-state Markov Chains for sero-status embedding within the hierarhical Bayesian framework for geostatistics. Results from a recent application of these techniques to data from the Cambodia malaria surveys will be presented by way of illustration.

5.5 INVITED - Spatial Statistics: Spatial Statistics for tropical disease control

Tuesday 6 September 2016 – 4.30pm-5.50pm

Advances in mapping malaria for elimination: fine resolution modelling of Plasmodium falciparum incidence

Victor Alegana, Peter Atkinson, Christopher Lourenço, Nick Ruktanonchai, Claudio Bosco, Arnaud Le Menach, Stark Katokele, Andrew Tatem
University of Southampton

The long-term goal of the global effort to tackle malaria is national and regional elimination and eventually eradication. Fine scale multi-temporal mapping in low malaria transmission settings remains a challenge and the World Health Organisation propose use of surveillance in elimination settings. Here, we show how malaria incidence can be modelled at a fine spatial and temporal resolution from health facility data to help focus surveillance and control to population not attending health facilities. Using Namibia as a case study, we predicted the incidence of malaria, via a Bayesian spatio-temporal model, at a fine spatial resolution from parasitologically confirmed malaria cases and incorporated metrics on healthcare use as well as measures of uncertainty associated with incidence predictions. We then combined the incidence estimates with population maps to estimate clinical burdens and show the benefits of such mapping to identifying areas and seasons that can be targeted for improved surveillance and interventions. Fine spatial resolution maps produced using this approach were then used to target resources to specific local populations, and to specific months of the season. This remote targeting can be especially effective where the population distribution is sparse and further surveillance can be limited to specific local areas.

5.5 INVITED - Spatial Statistics: Spatial Statistics for tropical disease control

Tuesday 6 September 2016 – 4.30pm-5.50pm

Geostatistical modelling of the relationship between microfilariae and antigenaemia prevalence of lymphatic filariasis infection

Emanuele Giorgi, Rachel Pullan, Jorge Cano
Lancaster University

Lymphatic filariasis (LF) is a mosquito-borne disease caused by parasitic roundworms of the genus *Wuchereria* and *Brugia*. Adult worms live and develop in the lymphatic system, causing fluid collection and swelling. LF is found throughout the tropics and sub-tropics, with Africa accounting for 30% of the global burden. The diagnosis of LF infection is mainly based on the detection of filarial offspring, called microfilariae (MF), and circulating antigens, secreted by adult worms, through an immunochromatographic rapid test (ICT). Despite its higher sensitivity and cost-effectiveness, a major caveat of ICT is its inability to measure the load of infection which ultimately determines the risk of transmission. We develop geostatistical methods in order to model the relationship between MF and ICT prevalence. More specifically, the mean number of worms and the basic reproduction number of LF are modelled as a pair of spatial Gaussian processes. We describe an application of this approach to model ICT prevalence data from Ethiopia and map MF prevalence based on an analytical expression which links the two diagnostics. We conclude by discussing the limitations of the developed methodology and the use of alternative empirical methods.

5.7 INVITED - Methods & Theory: Recent advances in Monte Carlo Methods

Tuesday 6 September 2016 – 4.30pm-5.50pm

Super-efficient sampling using Zig Zag Monte Carlo

Joris Bierkens, Paul Fearnhead, Gareth Roberts
University of Warwick

In recent work (Bierkens, Roberts 2015, http://arxiv.org/abs/1509.00302) we stumbled upon an elementary piecewise deterministic Markov process ('zig zag process') which can be extended to have a general (absolutely continuous) invariant probability distribution in R^n. We develop MCMC based on the zig zag process. The sample paths of the zig zag process can be efficiently simulated by rejection sampling of the switching times. It is possible to perform exact sub-sampling with the zig zag process, which makes this sampling method extremely promising for applications in big data and/or complex models. Somewhat mysteriously, using this method we can achieve a higher efficiency than independent draws from the correct distribution would allow!

Joint work with Paul Fearnhead and Gareth Roberts

5.7 INVITED - Methods & Theory: Recent advances in Monte Carlo Methods

Tuesday 6 September 2016 – 4.30pm-5.50pm

Rolling MCMC: Safe, Efficient Updating of Bayesian Models

Din-Houn Lau, Axel Gandy
Imperial College London

This presentation introduces the *rolling* MCMC (RMCMC) system that is able to update estimates from a sequence of probability distributions. The aim of the system is to quickly and efficiently produce estimates within a user-specified bound on the Monte Carlo error. The estimates are based upon weighted samples stored in a database. The stored samples are maintained such that the accuracy of the estimates and the quality of the samples are satisfied — in this sense, the system is self-monitoring and takes the appropriate course of action. This maintenance involves varying the number of samples and updating the weights. When required, new samples are generated by a Markov chain Monte Carlo algorithm — in this sense the system is efficient.

The performance of the system is demonstrated for predicting the end of season ranks of the English Football Premier League. Future work and directions to develop the RMCMC system are also discussed.

Joint work with Axel Gandy.

5.7 INVITED - Methods & Theory: Recent advances in Monte Carlo Methods

Tuesday 6 September 2016 – 4.30pm-5.50pm

Non-Parametric Control Functionals to Improve Monte Carlo Integration

Mark Girolami
University of Warwick

A class of estimators for Monte Carlo integration is proposed that leverages gradient information on the sampling distribution to improve statistical eciency. The novel contributions of this work are based on two important insights; (i) a trade-o between random sampling and deterministic approximation and (ii) a new gradient-based function space, building on recent work by Mira et al. (2013). The proposed estimators can be viewed as a non-parametric development of control variates. Unlike control variates, however, our estimator achieve super-root-n convergence, often requiring orders of magnitude fewer simulations to achieve a xed level of precision. Theoretical and empirical results are presented, the latter focusing on integration problems arising in hierarchical models and models based on non-linear ordinary dierential equations.

Plenary 2 – Read Paper

Tuesday 6 September 2016 – 4.30pm-5.50pm

Should we sample a time series more frequently? Decision support via multirate spectrum estimation

Guy P Nason and Ben Powell, *University of Bristol, UK*

Duncan Elliott*, Office for National Statistics, Newport, UK*

Paul A Smith, *University of Southamption, UK*

Suppose that we have a historical time series with samples taken at a slow rate, e.g. quarterly. The paper proposes a new method to answer the question: is it worth sampling the series at a faster rate, e.g. monthly? Our contention is that classical time series methods are designed to analyse a series at a single and given sampling rate with the consequence that analysts are not often encouraged to think carefully about what an appropriate sampling rate might be. To answer the sampling rate question we propose a novel Bayesian method that incorporates the historical series, cost information and small amounts of pilot data sampled at the faster rate. The heart of our method is a new Bayesian spectral estimation technique that is capable of coherently using data sampled at multiple rates and is demonstrated to have superior practical performance compared with alternatives. Additionally, we introduce a method for hindcasting historical data at the faster rate. A freeware R package, regspec, is available that implements our methods. We illustrate our work by using official statistics time series including the UK consumer price index and counts of UK residents travelling abroad, but our methods are general and apply to any situation where time series data are collected.

6.1 Contributed - Methods & Theory: Missing Data

Wednesday 7 September 9am – 10am

Efficient estimation in semiparametric regression with missing responses

Ursula U. Mueller
Texas A&M University

I will first review some of my results on efficient estimation in semiparametric regression with responses Y *missing at random* based on imputation. Then I will demonstrate that characteristics of the conditional distribution of Y given the covariates X can be estimated efficiently using complete case analysis, i.e. one can simply omit incomplete cases and work with an appropriate efficient estimator which remains efficient.

The *efficiency transfer* is a general result and holds true for all regression models for which the distribution of Y given X and the marginal distribution of X do not share common parameters. The derivation uses the *transfer principle* for obtaining limiting distributions of complete case statistics (for general missing data models) from corresponding results in the complete data model.

6.1 Contributed - Methods & Theory: Missing Data

Wednesday 7 September 9am – 10am

Justifying the Fully Conditional Specification imputation procedure

Finbarr Leacy, Ian White
Royal College of Surgeons in Ireland

Fully Conditional Specification (FCS) is an iterative imputation procedure requiring specification of a univariate imputation model for each partially observed variable conditional on all remaining fully and partially observed variables. While FCS has become a popular approach for performing multiple imputation under the missing at random assumption, exploration of the theoretical properties of the procedure has been limited to settings in which the set of univariate imputation models is compatible with some well-defined joint density for the partially observed variables given the fully observed variables. We review the available theoretical results [1-3], comment on their practical significance, and highlight directions for future research. We introduce new classes of compatible distributions to the multiple imputation literature through synthesis of existing results from the spatial and graphical modelling literatures.

[1] Liu J, Gelman A, Hill J, Su YS, Kropko J. On the stationary distribution of iterative imputations. Biometrika *2014; 101:155-173.*

[2] Hughes RA, White IR, Seaman SR, Carpenter JR, Tilling K, Sterne JAC. Joint modelling rationale for chained equations. BMC Med Res Methodol *2014; 14:28.*

[3] Zhu J, Raghunathan TE. Convergence properties of a sequential regression multiple imputation algorithm. J Am Stat Assoc *2015; 110:1112-1124.*

6.1 Contributed - Methods & Theory: Missing Data

Wednesday 7 September 9am – 10am

Outcome-sensitive Multiple Imputation: a Simulation Study

Evangelos Kontopantelis, Iain Buchan, Ian White
University of Manchester

Background

Multiple imputation is used to deal with missing data in healthcare studies. It is known that the outcome should be included in the imputation model, but not whether it should be imputed. Similarly no clear recommendations exist on: the utility of incorporating a secondary outcome in the model, if available; the level of protection offered when data are missing not-at-random; the implications of dataset size vs. extent of missingness.

Methods

We used realistic assumptions to generate thousands of datasets across a broad spectrum of contexts: three mechanisms of missingness (completely-at-random; at-random; not-at-random); varying extents of missingness (20-80% missing data); and different sample sizes (1,000 or 10,000 cases). For each context we quantified the performance of the common multiple imputation approaches, and a complete case analysis for comparison purposes, on median absolute bias, coverage and power.

Results

Overall, the best performer was the model that included and imputed the outcome but deleted cases where the outcome was imputed. However, this model was not the best in all scenarios and was marginally better at times, when compared to models that included the outcome in the multiple imputation model and either predicted or not. Including the outcome of interest and all available covariates was clearly beneficial, even when missingness was extensive. Incorporating a secondary outcome, strongly correlated with the outcome of interest, made very little difference. The dataset size and the extent of missingness affected performance, as expected. The multiple imputation methods protected less well against missingness not-at-random than other mechanisms.

Conclusions

We recommend to impute the outcome of interest and then remove cases where the outcome is imputed, especially for low and moderate levels of missingness. For very high levels of missingness, the higher power obtained when imputing the outcome (and not dropping cases) makes this approach more appealing.

6.2 Contributed - Medical Statistics: Meta-analysis

Wednesday 7 September 9am – 10am

Comparison of methods for measurement error correction: Regression calibration, multiple imputation and Bayesian methods

Christen Gray, Ruth Keogh, Jonathan Bartlett
LSHTM

When an exposure-outcome association is linear, measurement error in a continuous exposure leads to underestimation of the effect. Difficulty in ascertaining the true exposure is common in epidemiology. Frequently, no gold standard is available and information about the error is inferred through the use of repeated measures.

Our reference method, the most commonly used in epidemiology, is regression calibration (RC). Multiple imputation (MI) and Bayesian methods have been suggested as potential alternative methods for measurement error correction in epidemiology. Our objective is to compare the efficacy of these methods through simulations studies. Our motivating example is a person's underlying long-term blood pressure (BP) as a predictor of coronary heart disease; day to day fluctuations in BP result in random error which dilutes the naïve association.

MI has become widely used in the field of missing data. If a gold standard is available in a subset of study individuals, the underlying exposure can simply be treated as missing. When only repeated measures are available, further assumptions are required.

Full specification of the model structure and direct sampling from posteriors makes Bayesian analysis very flexible. Use of Markov chain Monte Carlo (MCMC) methods for sampling makes it computationally feasible. However, given the time-consuming nature of MCMC with large data sets, an alternate Bayesian method relying on integrated nested Laplace approximations (INLA) was included.

All methods were tested using simulated data varying sample size, effect size, and measurement error variance for linear regression and logistic regression. RC does not fully correct for bias when the effect size or measurement error variance is large and the sample size is small. MI performed similarly to RC. INLA, while faster, did not fully correct for bias when measurement error variance and effect size were large. Only Bayesian MCMC resulted in minimization of bias for all scenarios.

6.2 Contributed - Medical Statistics: Meta-analysis

Wednesday 7 September 9am – 10am

The albatross plot: a novel graphical tool for presenting results of diversely reported studies in a systematic review

Sean Harrison, Hayley Jones, Julian Higgins
School of Social and Community Medicine, University of Bristol

Meta-analyses combine the results of multiple studies of a common question. Approaches based on effect size estimates from each study are generally regarded as the most informative. However, these methods can only be used if comparable effect sizes can be computed from each study, and this may not be the case because of variation in how the studies were done or because of limitations in how their results were reported. Other methods must then be used to summarise the results of these studies. One possibility is a simple vote counting method, where studies are divided by statistical significance and direction to give an overall indication of the number of studies showing an association. Preferable to vote counting, meta-analysis of P values can by undertaken using Fisher's or Stouffer's method. These methods have important limitations however, due to the well-known pitfalls of P values and in particular their dependence on sample size: without sample size, a given P value could have any magnitude of effect.

We propose a novel plot that requires only a P value, a total sample size and a direction of effect from each study. Notably, the plot allows an approximate examination of underlying effect sizes and the potential to identify sources of heterogeneity across studies. This is achieved by drawing contours showing the range of effect sizes that might lead to each P value for given sample sizes, under simple study designs. These contours enhance the interpretability of the albatross plots, so named because the contours resemble the eponymous bird's wings. Examples of albatross plots using real data are presented, and their production and utility are discussed.

6.2 Contributed - Medical Statistics: Meta-analysis

Wednesday 7 September 9am – 10am

Individual participant data meta-analysis for external validation and recalibration of a prognostic model

Joie Ensor, Paul C. Lambert, Kym I.E. Snell, Thomas P.A. Debray, Karel G.M. Moons, Richard D. Riley
Keele University

Introduction: The availability of individual participant data (IPD) from multiple sources allows the external validation of a prognostic model across multiple settings and populations. When applying an existing prognostic model in a new population it is likely that it will suffer from some over or under fitting, potentially causing poor predictive performance. However, rather than discarding the model outright, it may be possible to modify components of the model to improve its performance using recalibration techniques. Here, we consider how IPD meta-analysis methods help identify and compare the best recalibration strategy, or whether a completely new model is warranted.

Methods: We examine four options for recalibrating an existing flexible parametric survival model in breast cancer across multiple centres and countries: (i) shifting the baseline hazard by a constant, (ii) re-estimation of the shape of the baseline hazard, (iii) adjustment of the linear predictor as a whole (calibration slope), and (iv) finally adjustment of individual predictor effects. We use IPD meta-analysis to examine calibration and discrimination performance across centres for each of the strategies.

Results: IPD meta-analysis reveals that re-estimation of the intercept gave the greatest improvement in calibration in new populations, with minor additional improvements seen after re-estimation of the baseline hazard. Scaling the linear predictor as a whole had little impact compared to adjustment of individual predictor effects. Heterogeneity in performance across centres was also reduced substantially when the intercept was re-estimated, and compared well in comparison to developing an entirely new model using all the IPD.

Conclusions: IPD meta-analysis methods allow different recalibration strategies to be compared when applying an existing model in new populations, and helps ascertain which components should be updated and when it is better to develop a completely new model.

6.3 Contributed - Social Statistics

Wednesday 7 September 9am – 10am

Evaluation of student performance through a multidimensional latent class IRT model with nonignorable missingness

Leonardo Grilli, Silvia Bacci, Francesco Bartolucci, Carla Rampichini
University of Florence

The paper focuses on the performance of university students, with reference to first-year compulsory courses. In particular, we consider the freshmen of the School of Economics at the University of Florence. In the Italian academic system, a student can enroll for an exam immediately after the end of the teaching period or can postpone it to any later examination session, so that the grade is missing until the exam is not attempted. We propose an approach for the ongoing evaluation of student proficiency accounting also for non-attempted exams. The approach is based on considering each exam as an item, so that responding to the item amounts to attempting the exam, and on an Item Response Theory model that includes two discrete latent variables corresponding to the student ability and the propensity to attempt the exam. In this way, we explicitly account for non-ignorable missing observations as the indicators of item response also contribute to measure the ability and then the model is of within-item multidimensional type. The two latent variables are assumed to have a discrete distribution defining latent classes of students that are homogenous in terms of ability and priority assigned to exams. The model also allows for individual covariates in the structural part. Estimation is performed by a specifically developed R package based on the Expectation-Maximization algorithm. The application to the data from the School of Economics at the University of Florence revealed that student decisions about attempting or postponing the exams depend on both the general ability and the preference for quantitative disciplines. The proposed approach allowed us to cluster students into four latent classes of ability, corresponding to widely different performances.

6.3 Contributed - Social Statistics

Wednesday 7 September 9am – 10am

Nurse effects in survey biomarkers

Alexandru Cernat, Joseph Sakshaug
University of Manchester

Surveys are extending the types of data they collect. An important new source is biological data such as blood or saliva collected from respondents. Nevertheless, this exciting new data source brings with it also a number of methodological issues. One of the most important ones is selection due to non-response and lack of consent to biological data collection. Our research aims to tackle this issue by looking at how nurses, respondents and areas influence 1) the likelihood that individuals will respond to the survey, 2) consent to biological data collection, and 3) the quality of the biological samples collected in two major British surveys: Understanding Society and English Longitudinal Study of Aging. This research will inform how nurses can have an impact on selection and data quality and how to model these effects as part of substantive research. It will also inform possible ways in which data collection could minimize nurse effects.

6.3 Contributed - Social Statistics

Wednesday 7 September 9am – 10am

Educational Qualifications Yields as Employment Risk: an Empirical Analysis on the Horizontal Inequality

Ivano Bison, Maria Michela Dickson, Giuseppe Espa, Flavio Santi
University of Trento

The analysis of horizontal inequalities in the educational yields has recently raised an increasing interest, as they explain a larger portion of the inequalities in the labour market, both in terms of wages and in terms of employment risk. This paper analyses the horizontal inequalities in terms of one-year employment risk for bachelor and master's graduates between 2008 and 2013 at the University of Trento (Italy). It is studied the employment situation of students one year after graduation, distinguishing between those who are unemployed, those whose job is consistent with their academic career, and those who work in a different sector. It is also considered the portion of graduates who are involved in an internship, as this seems to have became an important way of access to the Italian job market. In order to describe both the multiplicity of the employment situations and the hierarchical clustering of graduates with respect to their degree courses, data are modelled by means of multilevel multilogit models. The analysis is based both on administrative data of the University of Trento and on results of the periodic surveys on Italian graduates conducted by the Italian intercollegiate consortium Almalaurea; this allows us to assess the effect of the regularity of the educational path, the time of graduation, the final mark and other characteristic of students' academic careers on their employment situation one year after graduation.

6.4 Contributed - Trial Design

Wednesday 7 September 9am – 10am

Experiment Design and Analysis Using Pre-planned Orthogonal Contrasts

Colin Birch
Animal and Plant Health Agency

Scientists overuse multiple comparisons and multiple comparison adjustments in the analysis of their experiments. Even if multiple comparison adjustments correct the type I error in an experimental analysis, they increase the type II error, often to high levels. High type II error can be a major problem in experiments performed as regulatory checks, or where there is an ethical or economic requirement to design efficient experiments, for example when animals are used. Scientists also often find power analysis of experiments with multiple treatments challenging, so they potentially design oversized experiments by basing sample sizes on power analysis of paired comparisons, and are unaware of the benefits of more sophisticated designs.

This presentation will show that contrasts provide a long established but neglected tool for addressing these problems. Orthogonal contrasts can be used to avoid multiple comparison adjustments in most small experiments, at least up to five treatments. A major benefit of designing experiments for pre-planned orthogonal contrasts is that scientific judgement is required to decide and structure the comparisons an experiment aims to perform. However, design should be elementary for experiments with three or four treatments, for which few alternative designs are likely. An added benefit of analysis by orthogonal contrasts is its partitioning of sources of variation, which allows interpretation of the relative contributions of factors considered in an experiment. Power analysis of experiments with multiple treatments is simple and intuitive using orthogonal contrasts, allowing consideration of two or more alternative outcomes. Although rarely used to analyse experiments with unequal sample sizes, orthogonal contrasts can analyse quite complicated designs with unequal sample sizes reasonably easily, potentially providing a more transparent alternative to generic black box algorithms within software packages.

6.4 Contributed - Trial Design

Wednesday 7 September 9am – 10am

Demographic and other influences on the efficacy of water conservation measures on water consumption patterns in student residences.

Deirdre Toher, Chad Staddon, Karen Simpson
University of the West of England

A long-term study of the efficacy of water conservation measures is ongoing in student residences at the University of the West of England Bristol. Blocks of student residences have been divided into a control group, where no additional conservation measures have been put in place, and then into blocks which had a variety of water conservation measures installed. Each block consists of accommodation for between 66 and 84 students and is defined as having a single entrance point, with a total of 1908 bedrooms in the study.

Water consumption has been measured every 30 minutes since January 2014 in each of the 24 blocks. One of the original aims of the study was to investigate the long-term efficacy of water conservation measures by passive consumers – rather than consumers who had opted into water conservation schemes. Another aim was to determine if these interventions were more likely to fail and subsequently cause leaks. As the population of the student residences changes each year, we have been able to investigate the influence of demographic differences on the patterns of water consumption within the student accommodation.

Daily consumption figures are modelled to avoid some of the noise that is associated with measurements on such a fine timescale over a prolonged period of time. We have used generalized linear modelling with day of academic year, day-of-week, intervention status and demographic information as explanatory variables in order to try to understand the complex relationship between the demographic breakdown of occupants and the overall consumption patterns of water.

6.4 Contributed - Trial Design

Wednesday 7 September 9am – 10am

Challenges in designing and analysing trials in road safety

Caroline Wallbank, Kevin McRae-McKee
Transport Research Laboratory

Designing and analysing trials and surveys in road safety can be challenging. The road environment is a risky place to be, yet many of the trials require real-world data to be collected. Randomised control trials are uncommon and often studies are retrospective using data not collected specifically for the purposes of the trial; for example speed data collected by Automatic Traffic Counters in the road. As a result, the data may not be available in the most appropriate format for analysis.

A range of data collection techniques are frequently used; these include on-road monitoring of vehicles or drivers, attitudinal questionnaires, or data from simulated journeys. These may be subject to bias (for example, response bias in questionnaire data), small sample sizes (due to budgetary constraints) and missing data.

Observational studies of vehicles, often carried out using video data, can result in large sample sizes. However, this brings its own challenges. We may want to assess how representative the sample is of the general driving population, but little is known about the drivers themselves.

This presentation describes the experimental design of numerous on-road trials and surveys carried out by TRL over the past few years; discusses some of the challenges faced and describes how these were resolved to answer the research questions.

6.5 Contributed - Spatial Statistics: Uncertainty and correlation

Wednesday 7 September 9am – 10am

Dynamic chloropeths for representing the uncertainty in spatial estimates

Geoffrey Jones
Massey University

Spatial estimates, such as small-area estimates of deprivation measures, are often presented as a chloropeth, ie a map of the country or region with sub-regions coloured differently to represent the size of the estimates. Such a presentation tends to be taken as the exact truth by users, who may be basing important decisions such as aid allocation on the maps. However the estimates may have considerable uncertainty attached to them, so that ranking or pairwise comparisons of sub-regions are not as precise as the map may suggest. It is difficult to represent this uncertainty to the user. A separate chloropeth showing the standard errors does not adequately convey the message. One possibility, if the map is to be viewed on-screen, is to make the chloropeth dynamic, with the colours changing according to the uncertainty of the estimates. An example is presented showing small-area estimates of sub-divisional poverty rates in Bangladesh. Some of the difficulties and issues in producing a suitable dynamic map are discussed.

6.5 Contributed - Spatial Statistics: Uncertainty and correlation

Wednesday 7 September 9am – 10am

A Case-Study Comparison of Methods for Factor Analysis of Spatially Correlated Multivariate Responses

Samuel Oman, Bella Vakulenko-Lagun, Michael Zilberbrand
Hebrew University of Jerusalem

Many environmental and biostatistical applications involve spatially dependent vectors, observed at different sites in a sampling region. Examples are concentrations of different chemicals in topsoil samples, and percent cover of different species on line transects or in subplots. It is often useful to perform a principal component or factor-analytic type of analysis, in order to understand and model both the correlations within the vectors and the structure of their spatial correlation. A model commonly used in geostatistical applications is the Linear Model of Coregionalization (LMC). We shall first give a brief description of the LMC, together with three recently proposed alternatives. We shall then apply the four methods to analyze data on concentrations of major ions in water samples taken from springs in a carbonate mountain aquifer. The methods give quite different results, with those of the LMC being much more interpretable. We conclude with some possible explanations for this difference in interpretability.

6.5 Contributed - Spatial Statistics: Uncertainty and correlation

Wednesday 7 September 9am – 10am

On Surface Estimation under Gaussian Subordination

Sucharita Ghosh
Statistics Lab, Swiss Federal Research Institute WSL, Birmensdorf, Switzerland

Considering kernels that have absolutely integrable characteristic functions, we give a simple proof of the uniform convergence in probability of a kernel smoothed trend surface estimator in a nonparametric regression model where the errors are non-linear transformations of unobserved Gaussian random fields. We then note that in addition, a direct estimate of the variance of the surface estimator can be proposed. Eventually, this leads to a potential bandwidth selection algorithm which relies on the local stationarity type property of the errors. In the talk, applications to local estimates of distribution functions and spatial Gini coefficients are considered. An example from a global total column ozone data set (source: NASA) illustrates the ideas.

6.6 Contributed - Official Statistics: Crime and outliers

Wednesday 7 September 9am – 10am

Examples of recent innovations in official crime surveys: responding to the changing nature of crime, user needs and legislation.

Daria Gromyko
Home Office

The nature of crime is changing and so is its impact upon society. The approach to measuring crime needs to keep up with this shift, to ensure that official statistics remain relevant and continue to inform Government and the public of the true picture of different types of crime.

For example, the upcoming introduction of the Psychoactive Substances Act, which is due to be reviewed by Government in 2018, has created a need for new evidence around the misuse of psychoactive substances, and has led to innovations in the Crime Survey for England and Wales (CSEW), which has measured self-reported drug misuse in England and Wales since 1996.

Other recent innovations to the CSEW have been made in response to the growing focus on fraud and cybercrime. While this survey now provides detailed measures of these crimes against households, the Home Office's Commercial Victimisation Survey (CVS) has asked businesses about these crimes for a number of years, at the premises level. However, it is hypothesised that a premises-level survey may underestimate fraud and cybercrime against multi-premises organisations, which may only keep records of incidents at head office level. The Home Office (with Ipsos MORI) is working to develop a survey of head offices, to provide a more detailed and robust measure of fraud and cybercrime against businesses.

This talk will focus on the above examples of official surveys, to describe the challenges faced in producing robust and relevant national measures of crime.

6.6 Contributed - Official Statistics: Crime and outliers

Wednesday 7 September 9am – 10am

A comparison of automatic outlier detection methods for time series

Cathy Jones, Jennifer Davies, Duncan Elliott, Tariq Aziz, Atanaska Nikolova
Office for National Statistics

The Office for National Statistics (ONS) is the UK's independent producer of official statistics and National Statistical Institute. ONS produces thousands of seasonally adjusted time series and forecasts using X-13ARIMA-SEATS. The software uses ARIMA models to both clean time series prior to seasonal adjustment and for forecasting. It allows automatic model identification which determines the order of the ARIMA model and the inclusion of outliers and level shifts. The approach of automatic model identification is typically used where large numbers of time series need to be analysed. This paper provides an empirical evaluation of the automatic outlier identification procedures available in X-13ARIMA-SEATS compared to alternative methods from an indicator saturation approach and changepoint methods, considering the effect of the methods on both seasonal adjustment and forecasting.

7.1 INVITED - Methods & Theory: Recent advances in statistical learning

Wednesday 7 September 10.10am – 11.30am

Consistent Sequential Learning Algorithms for Highly Dependent Time Series

Azadeh Khaleghi
Lancaster University

One of the main challenges in statistical learning today is to make sense of complex sequential data produced daily, which typically represent interesting, unknown phenomena to be inferred. From a mathematical perspective, the sequential learning problem can be formulated as follows. Given a long, possibly growing, sequence of observations, the aim is to make inference on the stochastic mechanisms that produce the samples. This task is usually done under the assumption that the samples are i.i.d, or that their distribution belongs to a specific model class, e.g. HMMs. However, these assumptions sometimes undermine the possibly complex structure of the data and the potentially long-range inter-sample dependence. Moreover, since little is usually known about the nature of the data, it is important to address inference beyond parametric and modelling assumptions.

One way to incorporate inter-sample dependence is to assume that the process distributions that generate the data are stationary ergodic. Intuitively, stationarity means that the starting point at which the observations have been recorded does not bear any importance. Ergodicity means that asymptotically, any realisation of the process reveals complete information about its distribution. This paradigm has proved useful in a number of learning problems that involve dependent sequential data. At the same time, many natural problems already turn out to be impossible to solve in this framework. In this talk I will discuss the possibilities and limitations of sequential inference in the stationary ergodic framework.

7.1 INVITED - Methods & Theory: Recent advances in statistical learning

Wednesday 7 September 10.10am – 11.30am

How many communities are there?

Yi Yu, Diego Franco Saldana, Yang Feng
University of Cambridge

Stochastic blockmodels and variants thereof are among the most widely used approaches to community detection for social networks and relational data. A stochastic blockmodel partitions the nodes of a network into disjoint sets, called communities. The approach is inherently related to clustering with mixture models; and raises a similar model selection problem for the number of communities. The Bayesian information criterion (BIC) is a popular solution, however, for stochastic blockmodels, the conditional independence assumption given the communities of the endpoints among different edges is usually violated in practice. In this regard, we propose composite likelihood BIC (CL-BIC) to select the number of communities, and we show it is robust against possible misspecifications in the underlying stochastic block-model assumptions. We derive the requisite methodology and illustrate the approach using both simulated and real data. Supplementary materials containing the relevant computer code are available online.

7.1 INVITED - Methods & Theory: Recent advances in statistical learning

Wednesday 7 September 10.10am – 11.30am

Inference with Kernel Embeddings

Dino Sejdinovic
University of Oxford

Embeddings into reproducing kernel Hilbert spaces (RKHS) provide flexible representations of probability measures. They have been used to construct powerful nonparametric hypothesis tests and association measures and lead to a notion of Maximum Mean Discrepancy (MMD), a nonparametric distance between probability measures, popular in machine learning literature. I will overview recent developments within this framework: (1) a Bayesian method to estimate kernel embeddings leading to a new approach to select kernel hyperparameters and detect multiscale properties in the data, and (2) an application of kernel embeddings in the context of Approximate Bayesian Computation (ABC), where distribution regression and conditional distribution regression from the embeddings defined on simulated data to the parameter space can lead to "semi-automatic" construction of informative summary statistics.

References:

S. Flaxman, D. Sejdinovic, J. P. Cunningham, and S. Filippi, Bayesian Learning of Kernel Embeddings, in Uncertainty in Artificial Intelligence (UAI), 2016.
J. Mitrovic, D. Sejdinovic, and Y. W. Teh, DR-ABC: Approximate Bayesian Computation with Kernel-Based Distribution Regression, in International Conference on Machine Learning (ICML), 2016.

7.2 INVITED - Medical Statistics: Causality for medical statistics

Wednesday 7 September 10.10am – 11.30am

Causal inference with 0%, 50% and 100% invalid instruments

Jack Bowden
University of Bristol

Mendelian randomization (MR) (Davey-Smith and Ebrahim, 2003) uses multiple genetic variants that are hypothesised to satisfy the Instrumental Variable (IV) assumptions, for probing questions of causality in epidemiological research. The purpose of such an endeavour is clear: if a modifiable exposure can be shown to causally effect the severity (or risk of) disease, then interventions can be intelligently targeted to achieve a material benefit to public health.

Unfortunately, due to the increasing number of genetic variants that are now routinely included into MR analyses, it is more likely than ever that some do not meet the IV assumptions, due to exerting pleiotropic effects on the disease not through the exposure of interest. Furthermore, the traditional statistical approach to MR is known to fail even if only 1 of the included variants is an invalid IV. For this reason, we say it has a 0% breakdown level.

We discuss how this inherent weakness has fuelled the recent development of many new methods for MR, which trade off statistical efficiency for robustness to pleiotropy. Particular focus will be given to the technique of MR-Egger regression (Bowden et al, 2015) which can consistently estimate causal effects even when 100% of the variants are invalid IVs.

7.2 INVITED - Medical Statistics: Causality for medical statistics

Wednesday 7 September 10.10am – 11.30am

Efficacy and mechanisms evaluation using instrumental variables

Richard Emsley
University of Manchester

In many contexts, we want to go beyond investigating the effect of assigning treatments, and answer questions about whether the treatment works if you actually receive it or about how a treatment works. Methods for answering these questions, referred to as efficacy and mechanisms evaluation, have grown in recent years, particularly within the causal inference literature. However, the resulting estimates may not be causally valid without making strong assumptions about unmeasured confounding, even in randomised trials.

New statistical approaches have been proposed that acknowledges the likely presence of unmeasured confounding between post-treatment variables on the causal pathway between treatment and outcome. These methods include instrumental variables, structural mean models, and rank preserving structural models estimated by an iterative g-estimation procedure. We introduce these methods, and demonstrate the equivalence of the rank-preserving structural model and structural mean models estimated by g-estimation with an instrumental variables approach estimated by 2SLS. The instrumental variables approach can easily be implemented within standard statistical software. We illustrate a mathematical proof of the equivalence of the methods and use simulation studies to verify this, before analysing data from several psychological treatment trials using both methods.

7.2 INVITED - Medical Statistics: Causality for medical statistics

Wednesday 7 September 10.10am – 11.30am

Causality, Interaction and Observational Studies

David Cox
Nuffield College, Oxford

In 1965 W.G.Cochran read to the Society a very influential paper stressing in effect that, while issues of causality were reasonably well understood in the context of randomized experiments, there was much need for their discussion in connection with observational studies. Cochran mentions R.A.Fisher's gnomic comment to him that to strengthen causal interpretation one should ``make one's theories elaborate''. In proposing the vote of thanks Bradford Hill outlined his guidelines pointing towards causal interpretation, emphasizing that these were not necessary or sufficient criteria. Discussion of these issues has expanded over the years, including in particular two large recent books (Imbens and Rubin, 2015; Vanderweele, 2015) and much else. A misunderstanding about the various roles of interaction in connection with such investigations has resurfaced in recent years and will be one focus for the present paper. The role of randomization theory in addressing these issues will be outlined and some current challenges sketched.

7.3 INVITED - Social Statistics: Advances in segregation analysis

Wednesday 7 September 10.10am – 11.30am

Diversifying but not integrating: Entropy-based measures of local segregation in Philadelphia

Rory Kramer, Peter Kramer
Villanova University

With the substantial growth and spread of non-white immigrants in the United States in the past 40 years, scholars of segregation have struggled to adapt dichotomous metrics and indices designed to model two-group segregation for cities with large populations from more than two racial or ethnic groups. This paper argues that considering segregation as a social form of entropy is both axiomatically and theoretically beneficial and introduces a family of related measures that offer multiple means of analyzing a diversifying but segregated population at the local (neighborhood) level withing a greater (city/region) area. To illustrate, the paper introduces a case study of Philadelphia's level of segregation from 1990 to 2010. While Philadelphia has diversified, it remains racially segregated. Further, the entropy based measures show the growing importance of Hispanic segregation and a previously unidentified shift from segregation being visited upon the city's black population to a self-segregating white population that grows more segregated as its population share shrinks. The paper concludes by considering how incorporating separate measures of segregation *and* diversity into studies of residential racial patterns enhances our understanding of racial segregation levels and patterns in a multiracial context.

7.3 INVITED - Social Statistics: Advances in segregation analysis

Wednesday 7 September 10.10am – 11.30am

Developing a spatially disaggregated, multiscale index of dissimilarity: residential segregation between Asian and White British school children in England

Richard Harris
University of Bristol

A number of authors have promoted multilevel modelling as a way of measuring segregation simultaneously at multiple scales of a geographical hierarchy. In this way, micro-, meso and macro effects can be identified and estimated net of the effects at other levels of the model. This paper takes forward the approach by outlining a multilevel index of dissimilarity that allows both for localised estimates of difference and consideration to scale effects. Ways of measuring the contribution to the global index value of different places and at different levels of the geographic hierarchy are outlined. To demonstrate the method, a case study is made looking at the residential separation of Asian Commonwealth and White British pupils in 2011. The results suggest that school children are more residentially divided than is the Census population for the same year but that the majority of Asian Commonwealth and White British pupils were less segregated from each other in 2011 than they were in 2002. The spreading out of minority groups and the spatial contraction of the White British may be creating broad scale patterns of separation but this is taking place against a backdrop of decreasing segregation.

7.3 INVITED - Social Statistics: Advances in segregation analysis

Wednesday 7 September 10.10am – 11.30am

Frontiers in residential segregation

Nema Dean, Guanpeng Dong, Aneta Piekut, Gwilym Pryce
University of Glasgow

Social boundaries are potentially important features of the urban landscape. The frontiers between contrasting neighbourhoods are potential sources of conflict and also of enlightenment and interaction. But such boundaries are rarely truly “closed” – they may have a steep contrast with a particular neighbourhood in one direction, but blend more smoothly into neighbourhoods in other directions. This poses some formidable methodological challenges particularly if one seeks to compute inference for the existence of a boundary. This paper presents a method that addresses these issues using Bayesian spatial statistical methods. We also test whether neighbourhoods joined by social frontier tend to have higher rates of crime.

7.4 INVITED - Data Science: Tech Giants

Wednesday 7 September 10.10am – 11.30am

PSOne to PS4: The changing landscape of video game data and analysis

Aidan Fitzgerald
Sony Interactive Entertainment

Sales of the PS4 are currently at 40 million devices making it the fastest selling console in PlayStation's history. The connected nature of our homes and therefore devices provides us with a vast amount of information to analyse and understand the needs of our customers. The shift in philosophy within the games industry has been especially prevalent with the 7th generation of consoles. A shift from – physical media to digital, products to services/subscriptions, separated devices to interconnected networks. The storage, use and interpretation of this data at Sony Interactive Entertainment (SIE) will be discussed in this work. The role of a data scientist and the skills required within this industry will also be explored along with the contrast between PhD work versus business facing statistical work.

7.4 INVITED - Data Science: Tech Giants

Wednesday 7 September 10.10am – 11.30am

Modern Computing - Controlling the Data Explosion

Jonny Hancox
Intel

In all areas of our lives, data is being amassed at an unprecedented rate. The volume of this data is driving the demand for the next generation of computing and, in turn, each computing advance motivates scientists and engineers to tackle previously intractable problems.

Despite the considerable technical challenges involved, this vicious circle has delivered an explosive million-fold increase in supercomputing performance over the last 20 years. However, this additional computing power is of little use without the tools and techniques that can efficiently route the data through the computing medium and transform it into a format digestible by humans and, often, machines.

In his role as a software architect for Intel's Health & Life Sciences group, Jonny has witnessed many of the developments in today's big data analytics. In this session he will describe one or two examples of how silicon and statistics have been crafted into emerging branches of data science.

7.4 INVITED - Data Science: Tech Giants

Wednesday 7 September 10.10am – 11.30am

Big data and Machine Learning in today's analytical landscape

Bianca Furtuna
Microsoft

Terms such "big data", "machine learning", "data science" are extremely common when discussing today's data analytics landscape. These terms are constantly confused, mixed and interchanged to a point where their usage is not clearly defined anymore. In this session, I will be covering common architectures, usages and challenges of big data and machine learning in the tech industry as well as the benefits of using some of these technologies today.

7.5 INVITED - Geostatistics: Compositional Data Analysis

Wednesday 7 September 10.10am – 11.30am

The log-ratio approach to compositional data analysis

Juan José Egozcue
U. Politecnica de Catalunya

The definition of a composition is based on the principle of scale invariance: two arrays of positive values are compositionally equivalent if their components are proportional. For instance, a mineralogical composition is expressed equivalently in percentages or in parts per million. A composition is then an equivalence class, and a convenient representant is the projection in the simplex, i.e. a constant sum vector. The consequence of scale invariance is that relevant functions, e.g. statistics, should be scale invariant. Particularly, compositional information is contained in the ratios between components which are the simplest scale invariant functions.

A second principle is based on the fact that observed compositions are frequently obtained in sampling scenarios where different, but overlapping, sets of components are measured. Assuring coherence of results for the common sampled parts is critical. These principles led to the log-ratio approach in the eighties.

Around 2000 the simplex, as sample space of compositional data, was revealed to have its own interpretable geometry. Perturbation and powering are vector space operations which, combined with the Aitchison metrics, satisfy the properties of a Euclidean space. Consequently, any composition can be represented by Cartesian coordinates in the simplex. Isometric log-ratio (ilr) coordinates, which are real variables, are appropriate for any statistical analysis using standard statistical methods. In this framework, log-ratio transformation methods introduced in the eighties are particular choices of coordinates in the Aitchison geometry.

The Aitchison geometry of the simplex allows the definition of probability distributions on ilr-coordinates. For instance, the normal distribution in the simplex, satisfying a central limit theorem with respect to perturbation, is a distribution that outperforms the traditional Dirichlet class.

7.5 INVITED - Geostatistics: Compositional Data Analysis

Wednesday 7 September 10.10am – 11.30am

A compositional approach to investigating the relationship between environmental factors and health

Jennifer McKinley, Chloe Jackson, Ulrich Ofterdinger, Peter Atkinson, Damian Fogarty
Queen's University Belfast

The main focus of this research is to examine the potential relationship between environmental exposure to known nephrotoxins including arsenic, cadmium and lead and the potential health risk associated with the progressive dysfunction of the kidneys in renal impaired patients with Chronic Kidney Disease. The study uses a combination of datasets from the United Kingdom Renal Registry unknown aetiology subset and soil and stream geochemical regional survey data provided by the Tellus Survey, Geological Survey of Northern Ireland. Since geochemical data are observations that contain quantitatively expressed relative contributions of parts on a whole, a compositional data approach is presented that respects the relative nature of the geochemical data. The compositional multivariate nature of the soil and stream geochemical data is explored to aid in the analysis of interactions between elements. Using log ratio transformations, a data-driven and a knowledge-driven approach are explored to investigate the interaction between essential elements, which play a role as protecting mechanisms and those that increase the uptake of nephrotoxic elements as a result of similar absorption mechanisms within the body. Results from a compositional approach to Poisson regression analysis suggest a relationship between the presence of elevated arsenic in stream waters and impaired renal function of the kidneys.

7.5 INVITED - Geostatistics: Compositional Data Analysis

Wednesday 7 September 10.10am – 11.30am

Compositional geostatistics

Vera Pawlowsky-Glahn
University of Girona

Spurious correlation is known to be a problem in statistics since Pearson's early warnings in 1897. The same problems arise in spatial statistics: bias towards negative values and non-zero cross-covariances and cross-covariograms; singular matrices of intrinsic co-dispersion; co-kriged regionalised vectors of proportions that do not satisfy the constant sum constraint. A way out is to use the log-ratio approach introduced 1982 by John Aitchison, which has led to the Aitchison geometry of the sample space of compositional data. Within this framework, the spatial structure can be described in terms of direct variograms of each possible pairwise logratio; variation-variograms can be estimated even in case of missing components; they can be modelled with standard tools; both the data and the spatial structure model can be expressed in isometric logratio coordinates, and standard co-kriging techniques can be applied to obtain interpolated logratios. These can be back-transformed to compositions, delivering interpolated maps of each component that satisfy required constraints. Moreover, results do not depend on which logratio transformation was used for the computations. This approach and its potentialities is illustrated with a geochemical data set.

7.6 INVITED - Official Statistics: Innovation@ work in official statistics

Wednesday 7 September 10.10am – 11.30am

Deriving the Demographics of Twitter Users and exploring methods reduce bias in the data

Thomas Smith
Office for National Statistics

This paper investigates methods for deriving socio-demographic characteristics about Twitter users from information contained within their profiles. These variables are not explicitly available on Twitter, but may be derived using a mix of direct inference and machine learning techniques. This additional information should support better analysis of phenomena associated with Twitter data, but may also help in dealing with representivity issues within the data. It is established that Twitter users are not representative of the general population. A further line of enquiry is how enriched Twitter data could be combined with other data sources to produce estimates about student populations. The estimation methods explored incorporate the use of bespoke survey data to measure the usage of Twitter within the general population and then using this to adjust Twitter data to produce estimates that are more representative of the overall population. The research concludes that some socio-demographic characteristics (e.g. sex and language use) are easier derive directly than others (e.g. age). Where age cannot be inferred directly, a machine learning approach can be used to provide a very broad approximation. Although further refinements may be possible there are fundamental limitations with Twitter data and so good quality data for some variables may remain out of reach. However, the use of survey data to adjust for unrepresentative data produces plausible sub-national estimates of student populations. There are limitations associated with the sample size of the survey data, but this general approach to estimation looks promising and may prove to be important element in incorporating big data sources into official statistics.

7.6 INVITED - Official Statistics: Innovation@ work in official statistics

Wednesday 7 September 10.10am – 11.30am

Mixed-Mode Experiments an a Panel Survey – Changes in Response Behaviour

Tiina Orusild, Mikaela Jarnbert
Statistics Sweden

Over recent decades, response rates in interview surveys have decreased rapidly in Sweden as in many other countries. This has led to higher uncertainty in estimates as well as higher costs of data collection. One way to deal with this problem is to allow different response modes. In this paper we examine what happens with response behavior in a panel survey when web mode is introduced as an additional mode in a traditional single mode telephone interview survey.

Three experiments have been carried out in 2014 and 2015 mixing telephone interviews and web mode within a political opinion poll at Statistics Sweden. In all experiments the survey individuals have been randomly assigned into experimental and control groups. The experiments have shown positive results on increased response rates and no significant mode effects. Compared to the control group, the response rates are especially higher among earlier non contacts, but there are significant differences also among earlier refusals as well as among respondents from the previous survey round. The effect of offering web is also positive on new panels, where the respondents receive information on login by postal mail instead of e-mail.

Using three survey rounds in the overlap between single mode and mixed-mode we seek out to learn more about who answers, when they answer and how they answer. In the evaluation we use data from the surveys together with register data and para data from the data collection.

7.6 INVITED - Official Statistics: Innovation@ work in official statistics

Wednesday 7 September 10.10am – 11.30am

Data to Knowledge: Challenges and Opportunities for Official Statistics

Emanuele Baldacci
European Commission, DG ESTAT

Data are increasingly available at low production cost in today's world, as a result of technological innovation and changes in communication behaviors. With data becoming increasingly a commodity, the value of data is mostly related to their information and knowledge service content and the capabilities to use them for decisions.

Against this background, statistical organizations have to change their production chain. The European Statistical System Vision 2020 aims at innovation in four directions:

- multi-mode collection and use of multiple data sources. This entails a change from traditional stovepipe data collection modes and traditional sources (e.g., surveys) to other data sources encompassing administrative data and big data;
- data mesh-ups. Reliance on administrative sources in statistical production provides an opportunity for integrated data systems. Data need to be increasingly combined with new and eventually unstructured data sources (e.g., big data) through statistical models/methods and algorithms to produce on-demand statistical services.
- a new data "factory" to process in agile modalities multiple information queries. As a result production platforms will increasingly rely on interoperability and reuse properties to exchange tools and tasks across domains;
- data analytics services for "prosumers". Statistical organizations are increasingly moving in the direction of providing additional values to the statistical products offered to users, taking into account tailorization of services to put customers in the driver's sit in the consumption of statistical analytics.

Moving from data products to knowledge services in official statistics has great opportunities for increasing the relevance of statistical information services for data-driven decisions. This cannot be done in isolation by the statistical community and partnerships with the research and data science networks are critical to ensure success in the transformation agenda.

7.7 INVITED - Statistics & the Law: Statistics & the Law

Wednesday 7 September 10.10am – 11.30am

Modelling Activity Level Evidence using Chain Event Graphs

James Smith, Anjali Mazumder
University of Warwick

Bayesian Networks (BNs) have been widely and successfully used as a framework for integrating evidence and other scientific information. However when performing inferences that need to accomodate activity level evidence and hypotheses such methods can be rather restricitive since they depend on a level of structural symmetry concerning the competing hypotheses. In cases when this structural symmetry is not present we would recommend instead the use of Chain Event Graphs (CEGs). A CEG is a generalisation of a discrete BN. It is able to represent directly such assymmetries and shares nearly all the desirable properties enjoyed by a BN. It has alreay been demonstrated as a useful modelling tool in other domains. Here we illustrate both the representational power and effectiveness of the new framework for calculating the strength of evidence behind sometimes complex and heterogeneous competing hypotheses using the details of a typical assault case.

7.7 INVITED - Statistics & the Law: Statistics & the Law

Wednesday 7 September 10.10am – 11.30am

Uncertainty in likelihood ratios for forensic DNA evidence

Bruce Weir
University of Washington

Although there is wide agreement that forensic DNA evidence is best presented with a likelihood ratio, the probability of the evidence under the prosecution hypothesis divided by the probability of the evidence under an alternative hypothesis, there is less agreement about the degree of, or even the meaning of, uncertainty for such ratios. In the simple case of a single contributor to an item of evidence, the likelihood ratio reduces to the reciprocal of the conditional probability of an untyped person having the evidential DNA profile given that a suspect has been seen to have that profile. Population genetic theory exists for these conditional probabilities for pairs of people in the same family, or in the same population, or more distantly related. Conditional, or match, probabilities are not the same as the profile probabilities and they depend on the probabilities of the constituent alleles and on measures of allelic dependence. Part of our uncertainty about the appropriate numerical values to assign to match probabilities is that we do not know the actual values of allelic dependencies even though we may have appropriate predicted values. The probability that two half brothers with the same mother both have the same maternal allele is 0.5, for example, but they either do or do not have the same allele for any genetic marker. Some authors advocate incorporating the variance of actual allele sharing, 0.25 in the half-sib case, into the likelihood ratio calculation, whereas other authors advocate constructing intervals for likelihood ratios. Other sources of uncertainty include sampling to provide allele probabilities and even choice of population to sample. The challenge for the forensic scientist remains how best to convey the strength of DNA evidence in a way that reflects the many judgments needed to produce the likelihood ratio.

Plenary 3 – Barnett Lecture

Wednesday 7 September 12noon – 1pm

The Carbon Club: Measuring and mapping carbon dioxide from remote sensing satellite data

Noel Cressie
National Institute for Applied Statistics Research Australia (NIASRA), University of Wollongong, Australia

This talk is about environmental statistics for global remote sensing of atmospheric carbon dioxide, a leading greenhouse gas. I suggest the use of spatio-temporal statistics based on conditional-probability models to handle uncertainties in the data and the processes. Specifically, I describe a spatio-temporal data fusion (STDF) methodology based on reduced-dimension Kalman smoothing. Data are fused that are space-based, ground-based, or model-based, from several instruments or sources with potentially different footprints, measurement-error characteristics, and data coverages. Here, processes are viewed as time series of spatial random fields; then the spatio-temporal covariance functions are derived from a reduced-dimensional process of known spatial basis functions whose random coefficients evolve in time. I illustrate the methodology on atmospheric carbon dioxide measurements (in parts per million) from different carbon-counting instruments.

This talk is based on joint work with Hai Nguyen and Amy Braverman of NASA's Jet Propulsion Laboratory

8.1 Contributed - Methods & Theory: Longitudinal Data

Wednesday 7 September 2pm – 3pm

A time-scale for cohorts with long follow-up

Margaret Hurley
University of Central Lancashire

Transforming data to achieve conformity to a statistical distribution is a cornerstone of much data analysis but is rarely done for survival times. This is due in part to the utility of Cox's proportional hazards' model which allows us to fit models without concern for the underlying survival distribution. For cohort studies with long follow-up, in which participants join at varying ages over lengthy periods of time, there are issues of accounting for improving longevity with advancing time and of modelling the effect of differing ages at entry. The former is usually addressed by using a stratified Cox model but this can result in many strata. The latter is addressed often by sub-models for the effect of age which can be difficult to identify. If risk factors such as an environmental exposure are of primary interest then differential age at entry and chronological time are nuisance parameters in this context. This talk describes the reference relative time-scale which is a transformation of a participant's survival time obtained by integrating the hazard in a reference population between the times of entry and exit of the participant in the cohort (1). This new time scale has units of 'expected lifetimes', has several interesting interpretations and links together traditional epidemiological methods with contemporary approaches in survival analysis and reliability. The methodology has been applied successfully to data from the UK cotton workers' cohort and has shown that light smoking is relatively more harmful to women than to men (2).

1.Hurley M.A. (2015) A reference relative time-scale as an alternative to chronological age for cohorts with long follow-up. Emerg Themes Epidemio, 12:18. http://www.ete-online.com/content/12/1/18

2.Hurley MA. (2014) Light smoking at baseline predicts a higher mortality risk to women than to men; evidence from a cohort with long follow-up. BMC Public Health, 14: 95. DOI 10.1186/1471-2458-14-95

8.1 Contributed - Methods & Theory: Longitudinal Data

Wednesday 7 September 2pm – 3pm

Identifying disease progression using longitudinal discriminant analysis

David Hughes, Gabriela Czanner, Arnost Komarek, Marta Garcia-Finana
University of Liverpool

Clinical data is often collected over time for multiple biomarkers of different types (e.g. continuous, binary, counts). The aim of our study is to accurately predict patients' disease status (or progression) by combining this multivariate information.

Longitudinal Discriminant analysis (LoDA) aims to model the progression of longitudinal biomarkers in different disease groups. Following this, new patients can be classified based on the longitudinal history of their biomarker values.

In this talk we describe a multivariate generalised linear mixed model with a mixture of normal distributions assumed for the joint distribution of the random effects, in order to model the longitudinal profiles of biomarkers. We use these models in a longitudinal discriminant analysis with a dynamic allocation scheme to assign to each individual a probability of belonging to a specific prognostic group that is updated each time new information is available.

There are two main benefits to using our proposed models in LoDA. THe first is that markers of various types can be incorporated into a model which takes into account the correlation between the various markers. Secondly, the mixture distribution assumed for the random effects allows more flexible modelling of correlation between and within biomarkers.

We apply our approach to clinical data from patients with epilepsy in order to identify, as early as possible, those patients who will not achieve remission from seizures and show that (i) it achieves very good classification accuracy, (ii) it shows promising lead times, allowing predictions to be made earlier than is currently possible and (iii) waiting for more data to be available does not necessarily improve predictive accuracy.

.

8.1 Contributed - Methods & Theory: Longitudinal Data

Wednesday 7 September 2pm – 3pm

A joint semiparametric mixed model for longitudinal data involving time-varying covariates

Reza Drikvandi
Imperial College London

Longitudinal studies often produce data with both time-invariant and time-varying covariates. Similar to classical regression models, models for longitudinal data ignore the covariate process and treat all covariates as fixed variables, while in addition to the response variable, the time-varying covariates also change over time. Moreover, longitudinal models assume that the response variable measured at each follow-up time depends on the time-varying covariates measured at that time point only, but the response variable could also depend on the history of the previously measured time-varying covariates. We propose a joint mixed model for the response variable and the time-varying covariates which not only takes into account the covariate process for time-varying covariates but also allows the response variable to depend on the history of time-varying covariates. In our modelling approach, the association between the response and the time-varying covariates is taken into account through correlated random effects. We use P-spline functions (of time) to capture the evolutions of both the response and the time-varying covariates over time. Another advantage of our joint model is that it can also be used for situations where the response and the time-varying covariates for each subject at each follow-up are not measured at the same time. The proposed model is investigated theoretically and practically, and motivated by data from an AIDS cohort study in which HIV+ subjects have CD4 cell count and viral load measured at repeated visits before and after receiving treatment.

8.2 Contributed - Medical Statistics: Probability Weighting

Wednesday 7 September 2pm – 3pm

Propensity score methods in the estimation of treatment effect modification in observational studies

Antonia Marsden, Richard Emsley, Will Dixon, Graham Dunn
University of Manchester

Background: Treatment-moderator interactions are commonly evaluated to investigate if the effect of treatment varies between subgroups of patients. When evaluated in observational studies, adjustments have to be made for variables that confound the treatment-outcome relationship. However, it may be that the relationship between the confounders and the treatment and/or outcome varies across the moderator, and not accounting for this in a propensity score model could potentially produce biased estimates of the treatment-moderator interaction.

Aim: To assess the performance of propensity score methods for confounding adjustment in the estimation of treatment-moderator interaction effects under various scenarios regarding the moderator-confounder relationships.

Methods: Monte Carlo simulations were used to reflect situations in which a treatment effect moderator had varying degrees of influence on both the relationship between the confounders and treatment receipt, and between the confounders and the outcome. The performance of different propensity scores adjustment methods (direct adjustment, weighting and matching) in which the propensity score 1) did not account for any moderator-confounder interactions, 2) included moderator-confounder interactions, and 3) was estimated separately in each moderator subgroup, were assessed and compared. Extensions will be made to the evaluation of multiple treatment effect moderators simultaneously.

Results: Excluding moderator-confounder interactions in the propensity score generation gave valid treatment-moderator interaction effect estimates when the moderator influenced the confounder-outcome relationship, but not when the moderator influenced the confounder-treatment relationship. Adjustment of subgroup-specific generated propensity scores and cohort level propensity scores with interactions between the moderator and confounders via weighting and matching gave unbiased estimates of interactions in all considered scenarios when the confounders were uncorrelated.

Conclusion: Failure to account for interactions between moderators and confounders on treatment receipt leads to biased estimates of the treatment-moderator interaction effects in observational studies. Including interactions between the moderator and confounders in the propensity score can rectify this.

8.2 Contributed - Medical Statistics: Probability Weighting

Wednesday 7 September 2pm – 3pm

Improving prognostic accuracy with large, complex datasets from multiple sources and the merging process

Christopher Cheyne, David Hughes, Arnošt Komárek, Gabriela Czanner, Simon Harding, Marta García-Fiñana
The University of Liverpool

Larger and increasingly complex datasets are being used to more accurately predict a variety of future events, such as, an individual developing a disease or recovery from a disease following treatment. These datasets can be constructed using data from multiple sources (e.g. demographic, disease-specific, imaging, genetic data, etc.) which adds additional complexity. This results in more in-depth analyses being required. A range of statistical methods can be applied to these datasets to classify individuals into groups (e.g. disease, non-disease groups, etc.). However, as datasets become larger and more complex, methods with the capability of dealing with such datasets are not as well developed.

This talk will focus on the data linking procedure for obtaining a dataset using different types of data from multiple sources. The example given will involve data from people with diabetes who are screened annually for diabetic retinopathy. Diabetic retinopathy is one of the main causes of blindness in the UK and the world. The final, merged dataset includes static, demographic and dynamic data, treatment and medical data both from GPs, screening data from an ophthalmology screening service and procedural data from a hospital ophthalmology clinic. Additionally, a novel multivariate discriminant analysis approach will be described and applied to this dataset with the aim to identify individuals who are at high/low risk of developing diabetic retinopathy in order to tailor their screening intervals to differing lengths depending upon their risk.

8.2 Contributed - Medical Statistics: Probability Weighting

Wednesday 7 September 2pm – 3pm

Optimal probability weights for inference with constrained precision

Michele Santacatterina, Matteo Bottai
Unit of Biostatistics - Karolinska Institutet

Probability weights are used in many areas of research including complex survey designs, missing data analysis, and adjustment for confounding factors. They are useful analytic tools but often contain outlying values, which can cause statistical inefficiencies. This issue is frequently addressed by replacing large weights with smaller ones or by normalizing them through smoothing functions. For example, weights above the 90th centile of their sample distribution are replaced with the 90th centile itself. While these approaches are practical, they are also prone to yield biased inferences.

We present a method to obtain optimal weights that yield minimum-bias estimators among all estimators with specified precision. The optimal weights are the solution to a constrained optimization problem and minimize the Euclidean distance from target weights among all sets of weights that satisfy a constraint on the precision of the weighted estimator. The Lagrange multipliers and objective function from the optimization problem can help assess the trade-off between bias and precision of the weighted estimator.

In a simulation study, optimal weights performed better than trimmed weights with respect to bias and mean squared error of weighted least-squares regression coefficients. The study also showed that optimal weights often provided large gains in precision at the cost of small bias.

We illustrate the use of optimal weights in an analysis of the effect of timing of treatment initiation on long-term health outcome in patient infected by human immunodeficiency virus. We used data from a comprehensive register in Sweden. Our findings indicated that the age at the start of treatment was a relevant effect modifier, and correct timing of treatment initiation was more important in younger patients.

8.3 Contributed - Social Statistics: Applied

Wednesday 7 September 2pm – 3pm

What influences driver frustration? - An experimental study of factors associated with driver frustration and overtaking intentions

Neale Kinnear, Shaun Helman, Kevin McRae-McKee
Transport Research Laboratory

Driver frustration is an issue to which many people can relate. This project sought to gain a better understanding of the factors that influence the level of driver frustration, and to understand the association between the level of driver frustration and overtaking intentions. Driver frustration is thought to lead to a greater likelihood of a driver lowering their risk threshold in order make progress, potentially resulting in an unsafe passing manoeuvre.

The impact of a variety of factors (speed, platoon length, proportion of heavy goods vehicles and the presence of oncoming traffic) on self-reported driver frustration and overtaking intentions on single and dual carriageway roads was assessed using a five-way mixed analysis of variance. To do so, participants viewed pre-programmed simulated drives. The between-subjects factor was defined by splitting the sample into two distinct groups: those under time pressure and those not. A full factorial design was utilised for the remaining factors in which each participant received every possible combination of treatments. The results show that although there are clear effects of speed and platoon length on driver frustration and overtaking intentions, these are moderated by others variables and that increased frustration does not always lead to greater intention to overtake.

Data about driver attitudes was also collected via a questionnaire. Cluster analysis suggested that three distinct clusters were present within the data (high, medium and low risk) and further analysis showed that the level of frustration of higher risk drivers were more influenced by the previously mentioned factors than that of lower risk drivers.

8.3 Contributed - Social Statistics: Applied

Wednesday 7 September 2pm – 3pm

Investigating the impact of Alcohol Consumption on Health and well being in an Aging population (ELSA)

Rosemary McNiece, Peter Soan, Imogen Middleditch, Riza Momin
Kingston University

There is currently much interest in investigating health and well being in ageing populations of the developed world (Banks et al. 2012) and the English Longitudnal study of Aging (ELSA) provides a rich resource for examining the general health and behaviours of the older UK population. A particular area of interest is alcohol consumption and its the effect on aspects of health, cognitive function and social interaction in the older population. Several studies have shown that high levels of alcohol consumption are associated with decreased cognitive function in elderly people, while another study found that moderate alcohol consumption does not affect cognitive function in women and may in fact actually decrease the risk of cognitive decline (Meir et al 2005).

Our research uses data from ELSA to investigate patterns of alcohol consumption within the cohort and aims to examine associations between varying levels of alcohol consumption and aspects of health and lifestyle and to examine changes over time (2002 – 2012). Our findings have revealed that increased alcohol consumption has no negative impact on cognitive function in this cohort.

Further current investigation into associations between alcohol consumption and quality of life indicators, social interactions and health measures is ongoing with early indications giving interesting and surprising results.

8.3 Contributed - Social Statistics: Applied

Wednesday 7 September 2pm – 3pm

Modelling Euromillions sales

Rose Baker, David Forrest, Levi Perez
University of Salford

An analysis of Euromillions sales data from inception in 2004 to late 2015 is presented. The aim is to shed more light on the `compatriot-win effect' (Baker, Forrest and Perez, 2015), in which sales increase in a country following a win there, and also to study the effect of changing economic indicators such as unemployment rate on lottery sales. Do people buy fewer tickets when times are hard, or do they in desperation buy more tickets? These unique data allow this question to be answered, because the 9 countries have experienced differing economic conditions over the years, which means that a relation between sales and unemployment rate cannot be ascribed to secular trend. Both the issues addressed are relevant to planners (e.g. will a tax on lotteries yield more or less in a recession) as well as being of general interest.

This is (unbalanced) panel data, with 9 countries followed over 11 years, and its statistical analysis is not straightforward. We used fixed effects for countries, and a random effect for draw number (time). There is an additional modelling problem in that the compatriot-win effect term can act as a proxy for trend, so naive results would be quite misleading. Extracting meaningful answers to questions of interest from such complex data raises many statistical issues (as well as some computational problems). The model cannot be entirely correct, but we follow Box's principle that `all models are wrong, but some are useful'. We need however to demonstrate that the answers are meaningful, eg by `triangulation', using alternative methodology to check results. Graphical indicators of goodness of fit are also important.

Preliminary results suggest that rising unemployment does reduce lottery sales, and confirm the compatriot-win effect.

8.4 Contributed - Data Science: Changepoints and distributions

Wednesday 7 September 2pm – 3pm

Developments in Changepoint Analysis: Nonparametric and Parallelisation

Kaylea Haynes, Paul Fearnhead, Idris Eckley
Lancaster University

We are now able to record and store more data than ever before. Many applications require efficient ways to analyse the volume and variety of data available. For example one way to analyse these data sets is to detect changes in the distributional properties. Applications where changepoint detection is relevant are finance, environmental (such as oceanography and climatology), acoustics and genomics.

Changepoint detection has had a lot of attention in recent years with people trying to develop methods that can cope with much larger data sets as well as being applicable to different types of data. In this talk we will briefly look at two developments: a nonparametric approach that can be applied to real world data sets that don't have the nice distributional properties that would normally be required in changepoint analysis and a parallelisation technique that can be used when the volume of data is large.

In the first part of this talk I will look at a nonparametric algorithm that can be used to detect changes in time series where the underlying distribution is unknown. Our proposed method uses an exact optimisation approach to detect changepoints with an approximate cost function based on the empirical distribution of the data. I will show how this method can be used to detect changes in heart-rate data recorded using a Fitbit activity tracker.

The second half of this talk will look at how we can parallelise changepoint detection to improve computational cost. We look at ways to distribute data across multiple cores and then merge the results to find the overall set of changepoints. This can be very useful in practise, especially now computational power allows for much larger data to be recorded.

8.4 Contributed - Data Science: Changepoints and distributions

Wednesday 7 September 2pm – 3pm

Distinguishing between long memory and changepoint models: A spectral classification procedure

Rebecca Killick, Ben Norwood
Lancaster University

In recent years we have seen the rise of the use of changepoint and long memory processes for modelling time series data. A changepoint process is a nonstationary process where at one (or more) time points the statistical properties of the series change. In contrast a long memory process is a stationary process whose autocorrelations decay slowly and as such there is large dependence at large lags. On simply considering these definitions one would think it would be hard to confuse the two. However, if one plots the stationary autocorrelation function for a changepoint process it can present features of a long memory process (large dependence at high lags) – even when there is no dependence structure in the original changepoint process! It is important not to confuse these two process as the inference from data or forecasts will be drastically different for each.

The work presented here constructs a classification procedure to identify if, for a given time series, a changepoint or long memory process is more appropriate. We build our classifier using the time-varying spectrum and bootstrap training data from fitted long memory and changepoint processes. If our data do indeed arise from a changepoint model then using the time-varying spectrum we see clear step changes in structure. However, if our data arise from a long memory model then the time-varying spectrum is (approximately) constant over time as the series is stationary. We compare our classification procedure with the likelihood based hypothesis test from Yau & Davis (2012) with favourable results (whilst acknowledging that the two have different aims). To conclude we apply the approach to examples from finance and the environment.

Yau, Chun Yip and Davis, Richard A. (2012). Likelihood inference for discriminating between long-memory and change-point models. Journal of Time Series Analysis, 33(4):649–664.

8.6 Contributed - Official Statistics: Journeys through Data

Wednesday 7 September 2pm – 3pm

Telling stories with data- Recent advances in integrating data science to improve data dissemination and communication of official statistics at local level

Agnieszka Plywaczyk, Sean Hayes, Ruki Gul, Joanna Kacorzyk, Shanthan Golden
London Borough of Hounslow

The current school management information systems help measure educational performance and transform data into official and national statistics. These systems are intended to help make inferences about pupil characteristics and attainment data at a local level.

It is the quality of our local understanding of how our schools and pupils perform that enables us to make the right interventions. However, gaining a consistent understanding of the emerging issues and successes can actually be challenging to achieve, due to the increasing volume of data available to us and the high number of different statutory outputs as well as some of the limitations in the available software.

This presentation will provide an overview of how data on pupils' characteristics and achievements are transformed into understandable and reusable information resources, using various data science applications. We would like to share our experience of making data more accessible and easier to handle by improving the methods of data dissemination and communication, discuss recent data improvements and conclude with the benefits and pitfalls of the changes we processed.

The main areas that we would like to address include optimisation of data validation, consistent and transparent data management, interactive and reproducible analysis and reporting, real-time data dashboards, the automation of recurring tasks and comprehensive data support that assures quality outputs. We will present various examples to illustrate the implementation and application of these ideas in Hounslow Council.

Keywords used in the presentation include: MS Office, R, RStudio, Power BI, SQL, markdown, git, Github, data platforms, cloud computing, and digital by default.

8.6 Contributed - Official Statistics: Journeys through Data

Wednesday 7 September 2pm – 3pm

Improving comparability - demystifying official statistics through harmonisation

Suzanne Ellis, Charlie Wroth-Smith
Office for National Statistics

The Office for National Statistics (ONS) Harmonisation Team has been working on harmonising survey questions, concepts and definitions. The key driving factor is to deliver quality official statistics that better meet user needs. Greater harmonisation benefits users through facilitating the interpretation of statistics by supporting comparability across sources and time both within the UK and internationally. The ONS harmonisation vision is that all inputs, processing and outputs for the Census and surveys and all data from administrative records will be harmonised, so that users can compare data from different sources with confidence and can merge and match data more easily.

A new area of work is the harmonisation of business statistics. The main driver of this is EUROSTATs Framework Regulation Integrating Business Statistics (FRIBS). The regulation requires ONS to move to a harmonised set of variables by 2019. Alongside this, and in collaboration, ONS are harmonising business survey questions where possible on the Electronic Data Collection programme over the next four years.

The harmonisation of administrative data is also a new and challenging area of work and forms a large part of this harmonisation work programme. The harmonisation team will be researching questions/definitions on current administrative systems with the intention of developing some administrative data harmonised principles.

The ONS have a long working relationship with the UK Data Service and together plan to improve the Variable and Question Bank to coordinate with the developments in the business and administration harmonisation work.

The paper will focus on what the ONS has achieved to date with harmonisation, including the development of a harmonised question library, harmonised principles for surveys and also what remains to be done. It also outlines the benefits of harmonising and details the issues and challenges faced when attempting to harmonise.

8.6 Contributed - Official Statistics: Journeys through Data

Wednesday 7 September 2pm – 3pm

The Migrant Journey: a Home Office analysis of the paths that migrants take through the UK immigration system

Katie Fisher
Home Office

Understanding the processes that migrants go through when they decide to stay in the UK or switch their immigration status helps us to understand how the immigration system operates and indicates how changes in immigration rules have had an impact. Since 2010, the Home Office has published annually the Migrant Journey report, which analyses data on visas and extensions of stay in two complementary ways, providing a forward-view and a backward-view analysis of the way in which different types of non-European Economic Area migrants change their immigration status or achieve settlement in the UK. The forward-view analysis examines the immigration status of annual cohorts of migrants, at the end of each year following their arrival. The backward-view analysis looks at those who were ultimately granted settlement, to see which visas they used on their first arrival to the UK. The Migrant Journey: sixth report was published in February 2016, and is based on cohorts of migrants issued visas during 2004 to 2009, and those granted settlement during 2009 to 2014. Comparison across cohorts allows us to add to the migration evidence base by highlighting changes in migrant behaviour over time. Here we present an overview of the Migrant Journey report and highlight key findings from the latest data.

9.1 INVITED - Methods & Theory: Journal of the Royal Statistical Society Series B Editors' Invited Session

Wednesday 7 September 3.10pm – 4.30pm

Hypothesis Testing for Automated Community Detection in Networks

Purnamrita Sarkar, Peter Bickel
University of Texas, Austin

Community detection in networks is a key exploratory tool with applications in a diverse set of areas, ranging from finding communities in social and biological networks to identifying link farms in the World Wide Web. The problem of finding communities or clusters in a network has received much attention from statistics, physics and computer science. However, most clustering algorithms assume knowledge of the number of clusters k . We propose to automatically determine k in a graph generated from a Stochastic Blockmodel by using a hypothesis test of independent interest. Our main contribution is twofold; first, we theoretically establish the limiting distribution of the principal eigenvalue of the suitably centered and scaled adjacency matrix, and use that distribution for our test of the hypothesis that a random graph is of Erd˝os-R´enyi (noise) type. Secondly, we use this test to design a recursive bipartitioning algorithm, which naturally uncovers nested community structure. We demonstrate the empirical performance of our algorithm using simulations and quantifiable classification tasks on real world networks with ground truth.

9.1 INVITED - Methods & Theory: Journal of the Royal Statistical Society Series B Editors' Invited Session

Wednesday 7 September 3.10pm – 4.30pm

Generalized additive and index models with shape constraints

Yining Chen
London School of Economics and Political Science

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

9.1 INVITED - Methods & Theory: Journal of the Royal Statistical Society Series B Editors' Invited Session

Wednesday 7 September 3.10pm – 4.30pm

Truncated Linear Models for Functional Data

Giles Hooker, Peter Hall
Cornell University

A conventional linear model for functional data involves expressing a scalar response variable in terms of a weighted integral of an explanatory functional covariate in which the weighting function is a parameter to be estimated. However, in some problems the support of this weight is a proper and unknown subset of the function's domain and is a quantity of particular practical interest. Motivated by a real-data example involving particulate emissions, we develop methods for estimating the upper end of this support along with other parameters in the functional linear model. We introduce techniques for selecting tuning parameters; and we explore properties of our methodology using both simulation and the real-data example mentioned above. Additionally, we derive theoretical properties of the methodology, and discuss implications of the theory. Our theoretical arguments give particular emphasis to the problem of identifiability.

9.3 INVITED - Social Statistics: Organising Big Data's contribution to policy

Wednesday 7 September 3.10pm – 4.30pm

Organising big data’s contribution to policy

John Rigby
University of Manchester

Many major institutions with a public purpose are integrating data from a range of sources into single systems of Big Data to to try to answer questions about the efficiency, effectiveness and appropriateness of intervention. The session outlines two major attempts to implement such systems in the context of innovation policy. The session then engages experts from the area of policy and statistics to address both the potential and the limitations faced by those building and using such large assemblages of information in terms of the relevance of the data to actual policy problems, the reliability of data, aggregation, incommensurability and privacy and human rights.

9.6 INVITED - Official Statistics: Bringing official statistics to a wide audience

Wednesday 7 September 3.10pm – 4.30pm

Bringing official statistics to a wider audience – what needs to be done?

Nick Woodhill
Government Statistical Service

Official statistics are for the benefit of our society and our economy as a whole, not only in government policy-making and the evaluation of government performance. The 2015 ‘Better Statistics, Better Decisions’ strategy for UK statistics challenges the Government Statistical Service to find innovative ways of making statistics and analysis more accessible, engaging and easier to understand. It also challenges us to keep pace with a fast changing world.

This session will present different perspectives on the challenge of communicating official statistics and what needs to be done to meet the vision of ‘Better Statistics, Better Decisions’.

Julie Brown (Head of Profession for statistics at DfT) to introduce the work of the GSS Presentation and Dissemination Committee, its successes and challenges. High profile speakers from Full Fact and Financial Times (Alan Smith OBE, Data Visualisation Editor) will reflect on this work from a user perspective – highlighting opportunities and challenges to increase the helpfulness of official statistics. Audience participation will be an important part of the session and everyone will have the opportunity to contribute and shape the debate.

9.7 Professional Development: Becoming an expert witness

Wednesday 7 September 3.10pm – 4.30pm

Expert evidence, statements and liaison with lawyers

Colin Aitken
The University of Edinburgh

Experiences in the preparation of witness statements and of liaison with lawyers will be described. The presentation will cover the work involved from the initial approach by lawyers or police to the submission of a final report. Little will be said about the presentation of evidence in court as the speaker has attended court only very occasionally and given evidence only very, very occasionally.

9.7 Professional Development: Becoming an expert witness

Wednesday 7 September 3.10pm – 4.30pm

My experiences as an expert witness

David Balding
UCL

I was an expert witness in scores of criminal, and a few civil, cases in the UK between 1993 and 2014. These were mostly related to the evaluation of DNA profile evidence for identification, including statistical and population genetics aspects. I have also advised on the estimation methodology for street value of drugs, inheritance disputes (involving DNA evidence), an experimental design patent in an IPR case and another design issue relating to sampling to assess construction quality (I may remember a few other cases between now and the meeting). I had no formal training at any point, but learned on the job. I had many good experiences and felt able in many cases to provide important assitance to the court, but also much frustration at the tedious delays and the deliberate tactics of obfuscation and time-wasting that are sometimes used. I learned that lawyers are often good "students": they tend to be smart but ignorant of most stats-related issues. It is frustrating that the England and Wales court of Appeal has appeared determined in recent judgments to limit the potential uses of statistics in criminal trials, but nevertheless there is still ample scope for statisticians to assist the court process. I will review my experiences and try to identify advice for the newcomer statistician.

9.7 Professional Development: Becoming an expert witness

Wednesday 7 September 3.10pm – 4.30pm

Becoming an expert witness

Tim Clayton, Roberto Puch-Solis
Forensics & Security Division, LGC

Expert evidence plays a pivotal role in litigation featuring in cases of personal Injury, contractual disputes, intellectual property, construction and defective goods. In this session we will focus on the Criminal Law of England and Wales. Forensic science forms an integral part in the prosecution of many of the more serious criminal offences and the criminal courts rely heavily on the testimony of experts to assist the finders of fact. The expert's opinions are likely to command great respect and to be considered authoritative and as such the finders of fact will probably attach great weight to their opinions. Unsurprisingly therefore, the reliability and integrity of expert evidence has been considered many times over the years by the Judiciary and other stakeholders. Today, there is a body of law and many supplementary rules governing the ethical responsibilities and conduct of experts and regulating the process and procedures behind the provision of expert evidence. These restrictions apply irrespective of whether the expert is making a 'one-off' appearance in a specific case or carries out those functions for as part of their occupation. In this session, we will summarise the requirements for the provision expert evidence and the participants will be directed towards the various resources that exist.

Plenary 4 – Campion (President's Invited Lecture)

Wednesday 7 September 5pm – 6pm

Statistical insights into the West African Ebola outbreak

Christl Donnelly
Imperial College London - on behalf of the WHO Ebola Response Team

Between December 2013 and April 2016, the largest epidemic of Ebola virus disease (EVD) to date caused more than 28,000 cases and more than 11,000 deaths in Guinea, Liberia, and Sierra Leone. Thorough analyses of these data have provided new insights into the epidemiology of EVD. Epidemiological analyses conducted during the West African epidemic have shown that large epidemics of EVD are preventable — a rapid response using classical approaches of EVD control can interrupt transmission and restrict the size of outbreaks, even in densely populated cities.

10.1 Contributed - Methods & Theory: Markov Chains

Thursday 8 September 9am – 10am

MCMC for weakly identifiable targets from matrix functions

Thomas House
University of Manchester

In many application areas (particularly observational sciences such as the study of epidemics) posterior distributions are very far from multivariate normal due to curvature, but relatively low-dimensional.

A natural approach to such a problem it to explore the posterior distribution via a Markov-chain Monte Carlo (MCMC) algorithm that adapts to local correlations in the posterior distribution, since these vary across the parameter space. There has already been significant success in use of geometric concepts in this context.

This talk will present a novel approach to local adaptation of MCMC based on matrix functions rather than geometric concepts that can be shown to out-perform existing algorithms in some natural asymptotic limits, as well as applications of this family of algorithms.

A new method for estimation of the marginal likelihood from MCMC output that is well-adapted to curved posteriors will also be presented, as well as asymptotic results bounding the error of this procedure.

10.1 Contributed - Methods & Theory: Markov Chains

Thursday 8 September 9am – 10am

Visualization of distance measures implied by forecast evaluation criteria

Robert Kunst
Institute for Advanced Studies

Traditional moment-based measures of predictive accuracy, such as the mean squared error (MSE) and the mean absolute error (MAE), assess the precision of forecasts in the framework of widely accepted metric spaces. Many researchers, however, pursue more complex targets, such as the mean absolute percentage error (MAPE), often motivated by an attempt to reduce the influence of scaling. Recently, it has been pointed out that the apparent advantage of scale independence comes at a price, as these criteria may be hampered by the non-existence of moments.

We argue that, additionally, most of these measures are characterized by asymmetry and by non-convexity of the implied neighborhoods. Asymmetry means that moving the actual closer to the forecast has a quite different effect from moving the forecast. For some criteria, even paradox effects can be generated, such as a deterioration of accuracy as the actual approaches the forecast. Non-convexity means that a linear combination of predictions may become a bad predictor for a linear combination of actuals. We illustrate all effects using contour plots. The summary message is a warning against the careless usage of relative asymmetric criteria in forecast evaluation.

10.1 Contributed - Methods & Theory: Markov Chains

Thursday 8 September 9am – 10am

The Computation of Mean First Passage Times for Markov Chains

Jeffrey Hunter
Auckland University of Technology

The presentation focuses on a comparison of a variety of computational techniques for finding the mean first passage times in Markov chains. The presenter has recently published a new accurate computational technique (Special Matrices, 2016) similar to that developed by Kohlas (Zeit. fur Oper. Res., 1986) but based on an extension of the Grassmann, Taksar, Heyman (GTH) algorithm (Oper. Res., 1985) for finding stationary distributions of Markov chains. In addition the presenter has recently developed a variety of new perturbation techniques for finding mean first passage times in Markov chains (ArXiv.org, 2016). These procedures are compared with other well known techniques, such as the standard matrix technique (Kemeny and Snell, 1960) and some simple generalized matrix inverse techniques developed by the presenter (Asia Pacific J. Oper. Res., 2007). Matlab computations are used to make comparisons using some test problems used in the literature for comparing computational techniques for stationary distributions.

10.2 Contributed - Medical Statistics: Informative observation

Thursday 8 September 9am – 10am

Application of Marginal Structural Models to unbalanced logitudinal health data (Clinical Cohort): A simulation study

Edmore Chamapiwa, David Reeves, Darren Ashcroft, Evangelos Kontopantelis
University of Manchester

Background:

Marginal Structural Models (MSMs), a class of structural causal models, are being increasingly used in the analysis of complex longitudinal health data because of their ability to give unbiased effect estimates of a time-varying treatment in the presence of time-varying confounding/mediating covariates. However, MSMs have shown good performance to settings where observations occur at regularly separated time points for all patients, whereas in “real-life” health record data, different patients are commonly seen and measured at different and irregular time points. In addition, the frequency with which a patient is seen may well be related to their health outcomes. The impact of irregular, but more realistic, data on the performance of MSMs is unknown.

Objective:

To evaluate the performance in effect estimation of inverse-probability-weighted MSMs in unbalanced longitudinal health data (clinical cohort)

Methods:

A simulation study was conducted to compare treatment effect estimates from inverse-probability-weighted MSM, adjusted and unadjusted regression models. Irregular longitudinal data was generated by sampling time between consecutive visits for an individual from an inverse Gaussian distribution. Treatment at each observation time was sampled from a Bernoulli distribution with likelihood of getting treated dependent on the confounder level, and confounder values were sampled from a Bernoulli distribution. Continuous outcome values were simulated from a Normal distribution. Data simulation and analysis were conducted in R.

Results:

This simulation study showed that inverse-probability-weighted MSMs outperform stratification based estimation methods when longitudinal data is unbalanced and when confounders and treatments are time-varying.

10.2 Contributed - Medical Statistics: Informative observation

Thursday 8 September 9am – 10am

Extensions in Robust Joint Modelling

Lisa McCrink
Queen's University Belfast

Over the last two decades, there has been an increasing amount of renal research on the effect of anaemia on the survival of patients undergoing haemodialysis. One biomarker of interest, an individual's haemoglobin levels, has been shown previously in renal research, through cross-sectional techniques, to have a detrimental impact on survival when it fluctuates over time.

Building upon this research, such an association will be further investigated through the extension of robust joint modelling methodology. Joint modelling techniques are becoming more commonly seen within current literature due to the obvious benefits of being able to simulataneously analyse patients' repeated measurements and the influence these have on the time-to-event process. Recent work has investigated and verified the need to use more robust approaches in the field of joint modelling to accommodate longitudinal outliers through t-distributional assumptions. Robust joint models improve the accuracy and precision of parameter estimation in the presence of longitudinal outliers.

This research will extend current robust joint modelling techniques through the incorporation of the cumulative effect of changing haemoglobin levels over time within the survival submodel. In doing so, the impact on survival of changing haemoglobin levels over time can be captured. The methodology presented will be illustrated using data obtained from the Northern Ireland Renal Information Service on 1,340 haemodiaylsis patients over a ten year period commencing in 2001.

10.2 Contributed - Medical Statistics: Informative observation

Thursday 8 September 9am – 10am

Time dependent ROC for disease incidence using multivariate markers under informative censoring

Cuiling Wang
Albert Einstein College of Medicine

Evaluating the accuracy of multivariate markers in diagnosing incident disease using time dependent receiver operating characteristic curve (ROC) analysis has received increasing attention. In many longitudinal studies, censoring of the disease status and longitudinal markers can be informative or non-random, such as that due to death or drop out process that is related to disease and longitudinal markers, which is commonly seen in aging studies. Based on a framework of illness-death model to handle censoring of disease due to death, we evaluate two approaches to further handle the non-random drop out. In one approach, we use auxiliary information to account for the non-random missing data process. In another approach, we perform a sensitivity analysis to model the non-random drop out process and evaluate its effect on the ROC estimation. Simulation studies are performed to evaluate the performance of the proposed methods. The methods are applied to an example from the Einstein Aging Study (EAS).

10.3 Contributed - Social Statistics: Global Development

Thursday 8 September 9am – 10am

Lot Quality Assurance Sampling for improving health systems in the developing world: a decision-making tool to empower health officers and inform health policy

Caroline Jeffery
Liverpool School of Tropical Medicine

Objectives

Lot-quality assurance sampling (LQAS) is a classification method, developed in the 1920s for industrial quality control. In 1991, a WHO consultation on epidemiological and statistical methods for rapid health assessments recommended that LQAS be developed further to monitor health programmes. LQAS is used to manage health services performance quickly and relatively inexpensively in a defined geographical area. We review the statistical underpinnings of LQAS and methodological extensions, presenting recent applications in health in the developing world.

Methods

Standard LQAS methodology is a two-stage sampling approach defined in a catchment area (CA), stratified by supervision areas (SA). Communities are selected in each SA with probability-proportional-to-size; typically one respondent, sampled randomly in each community, is interviewed with a structured questionnaire. LQAS health surveys traditionally measure binary outcomes, classifying SA-level coverage indicators as having reached a predefined target. Classification is based on a decision rule, determined from binomial or hypergeometric distributions. Data from multiple SAs is aggregated to provide CA-level coverage estimates with confidence interval.

Example and Conclusion

During 2003-2015, LQAS household surveys were completed in up to 65 Ugandan districts to monitor health indicators. LQAS was rolled out as a national health sector monitoring system in 2009 and the data merged into one super-database, permitting cross-time and cross-space epidemiological studies to take place as secondary data analysis. One study looked at factors associated with facility-based delivery (FBD) adjusting for multiple factors simultaneously, spatial heterogeneity, and time trends. The statistical model formulated a nascent early warning system to identify districts expected to have low prevalence of FBD in the immediate future. LQAS is an attractive tool for evaluating health services. The scaling up of LQAS in the developing world provides numerous opportunities to design and conduct complex statistical analyses and evaluations to inform health policy and formalise our understanding of health systems.

10.3 Contributed - Social Statistics: Global Development

Thursday 8 September 9am – 10am

The social context of youth's use of modern contraceptives in Kenya: a multilevel analysis

Elsie Akwara
University of Southampton

Clear improvements to youth sexual and reproductive health in Kenya have been made in recent years; however, progress has been inconsistent or slow in the last decade. Majority of the youth in Kenya face challenges of weak education and health systems, limited access to sexual and reproductive health and a dearth of jobs or income- earning opportunities. Fertility rates among this cohort remain high and modern contraceptive uptake has been minimal. This study builds on previous studies by advancing existing knowledge beyond understanding of individual and household determinant of contraceptive uptake.

This study examines the influence of both individual, household and community variables influencing modern contraceptive use among youth in Kenya within the Millennium Development Goal (MGD) era, using the 2003 and 2014 Kenya Demographic and Health Surveys (KDHS). Univariate, bivariate and multilevel binary logistic regression analysis was performed using generalized linear latent and mixed models (GLLAMM).

The odds of contraceptive use were higher for youth with one or more living children (Odds Ratio (OR) 4.0 and 5.8 for those with 1-2 children and 3 or more children respectively). Being from the middle (OR, 1.4) and rich (OR, 1.2) wealth status, having secondary or higher education (OR, 1.3) and having medium (OR, 1.7) or high (OR, 1.9) access to media also increased the odds of using modern contraceptives. At the community level, the odds of modern contraceptive use decreased with an increase in the average number of children desired by women (OR, 0.8). This was the only community level variable that showed a significant predictor of modern contraceptive use among youth women in Kenya.

10.3 Contributed - Social Statistics: Global Development

Thursday 8 September 9am – 10am

Reaping digital dividends or emerging digital divide?

Gindo Tampubolon, S Sujarwoto, S Sukardi, Ahmad Nasution, Ardi Adji
University of Manchester

The 2016 World Bank development report extolls the dividends accruing to the poor when they were given access to the Internet and social media. Meanwhile, a recent report from Indonesia notes that spatial disparity across the 17,000 islands archipelago is reinforced by uneven access to the Internet. Access to the Internet and social media is likely to exhibit complex patterns of associations and consequences that can be understood both as dividends accruing to some groups and as a disparity limiting others.

We examine the most recent inter-censal survey of Indonesia from 2015 to bear on the questions of whether access to the internet and social media benefits the poor across the archipelago. Some 660,000 households (about 2.3 million individuals) were surveyed, eliciting information about their use of the Internet and social media including facebook, twitter, whatsapp, and instagram. Generalised linear mixed model was applied to the multilevel data of individuals nested within households and districts; the 511 highest level units were politically responsible for investment in public infrastructure. Variance partitioning suggests variations between districts separate from thosebetween individuals and households.

The analysis found inequality in access to the Internet and social media along social positions (assets, education and occupation), and spatial disparity across districts to be resistant to the spread of telecommunication infrastructure and the reduction in the prices of smart devices. These are the basis for a further exploration of the notion of unhealthy digital divide and a discussion on the impact of internet access on securing better quality occupations.

10.5 Contributed - Environment Statistics: Sea

Thursday 8 September 9am – 10am

Multi-dimensional Bayesian adaptive regression splines for marginal extreme value analysis

David Randell, Emma Ross, Philip Jonathan
Shell

Extreme values of many responses of environmental interest, such as ocean storm severity, vary systematically with covariates. Reliable design of coastal and marine infrastructure needs to allow for non-stationarity of the ocean environment with respect to at least four covariates (e.g. longitude, latitude, direction and season). Flexible modelling of the effects of multiple covariates can be achieved using a spline parameterisation, typically on a gridded multidimensional covariate domain. As covariate dimension increases, computational burden grows despite slick linear algebra methods such as GLAMs. Bayesian adaptive (or free knot) spline formulations, requiring a lower number of spline knots and hence parameters for typical applications, are therefore an attractive option.

We estimate the distribution of ocean storm peak significant wave height at a northern North Sea location using a two-part Weibull - generalised Pareto model, the parameters of which vary smoothly with respect to direction and season and are parameterised using adaptive splines. Parameter estimation is achieved using mMALA reversible jump MCMC. We compare inferences, in terms of statistical and computational efficiency, with those obtained using gridded penalised B-spline parameterisations. We show that both approaches provide estimates for return values of simlar quality, but that adaptive spline solutions are computationally more efficient, especially for higher-dimensional covariates.

10.5 Contributed - Environment Statistics: Sea

Thursday 8 September 9am – 10am

Extremal spatial dependence of North Sea storm wave environments

Monika Kereszturi, Mirrelijn van Nee, Emma Ross, David Randell, Jonathan Tawn, Philip Jonathan
Lancaster University

In environmental extreme value modelling, interest sometimes lies in the probability of joint occurrences of rare events. For example, we might wish to estimate the chance that two or more locations in the ocean experience severe storm conditions at the same time, and hence the risk to multiple marine structures from a storm event. To achieve this reliably, effective tools to characterise joint tail behaviours are required, are are fit-for-purpose spatial extremes models at appropriate levels of sophistication.

There are two different classes of joint tail behaviour that have very different implications: asymptotic independence suggesting that extreme events are unlikely to occur together, and asymptotic dependence implying that extreme events can occur simultaneously. It is vital to have good diagnostics to identify the appropriate dependence class. If variables are asymptotically independent, incorrectly assuming an asymptotically dependent model can lead to overestimation of the joint risk of extreme events, and hence to higher than necessary design costs of offshore structures. We develop improved diagnostics for differentiating between these two classes, which leads to increased confidence in model selection.

Application of the diagnostics to storm severities at pairs of locations throughout the North Sea suggests that tail dependence changes not only with distance between spatial locations, but also with the direction of storm propogation. Motivated by this, for the same application, we also estimate a number of max-stable spatial extreme value models, and consider evidence for directional and spatial non-stationarity of model parameters.

10.5 Contributed - Environment Statistics: Sea

Thursday 8 September 9am – 10am

An evolutionary spectra approach to model land/ocean nonstationarities

Stefano Castruccio, Joseph Guinness
Newcastle University

We introduce a nonstationary spatio-temporal model for gridded data on the sphere. The model specifies a computationally convenient covariance structure that depends on heterogeneous geography. Widely used statistical models on a spherical domain are nonstationary for different latitudes, but stationary at the same latitude (axial symmetry). This assumption has been acknowledged to be too restrictive for quantities such as surface temperature, whose statistical behavior is influenced by large scale geographical descriptors such as land and ocean. We propose an evolutionary spectrum approach that is able to account for different regimes across the Earth's geography, and results in a more general and flexible class of models that vastly outperforms axially symmetric models and captures longitudinal patterns that would otherwise be assumed constant. The model can be estimated with a multi-step conditional likelihood approximation that preserves the nonstationary features while allowing for easily distributed computations: we show how the model can be fit to more than 20 million data points in less than one day on a state-of-the-art workstation. The resulting estimates from the statistical model can be regarded as a synthetic description (i.e. a compression) of the space-time characteristics of an entire initial condition ensemble.

10.6 Contributed - Sport Statistics: Detection and Prediction

Thursday 8 September 9am – 10am

In-play forecasting of win probability in One-Day International cricket: A dynamic logistic regression model

Muhammad Asif
University of Malakand

We present a model for forecasting the outcomes of One-Day International cricket matches whilst the game is in progress. Our ‘in-play’ model is dynamic, in the sense that the parameters of the underlying logistic regression model are allowed to evolve smoothly as the match progresses. The use of this dynamic logistic regression approach reduces the number of parameters required dramatically, produces stable and intuitive forecast probabilities, and has a minimal effect on the explanatory power. Cross-validation techniques are used to identify the variables to be included in the model. We demonstrate the use of our model using two matches as examples, and compare the match result probabilities generated using our model with those from the betting market. The forecasts are similar quantitatively, a result that we take to be evidence that our modelling approach is appropriate.

10.6 Contributed - Sport Statistics: Detection and Prediction

Thursday 8 September 9am – 10am

Statistical Methodology for Age-Adjustment of the GH2000 Score Detecting Growth Hormone Misuse

Dankmar Boehning, Richard Holt, Peter Sönsken, David Cowan, Nish Guha
Southampton Statistical Sciences Research Institute

Background. The GH2000 score has been developed as a powerful and unique technique for the detection of growth hormone misuse by sportsmen and women. With the collection and establishment of an increasingly large data base it has become apparent that the score shows a positive age effect in the male athlete population, which could potentially place older male athletes at a disadvantage.

Methods. We have used results from residual analysis of the general linear model to show that the residual of the GH2000 score when regressed on the mean-age centered age is the right way to proceed to correct this bias. As six GH2000 scores are possible depending on the assays used for determining IGF-I and P-III-NP, methodology had to be explored for including six different age effects into a unique residual. Meta-analytic techniques have been utilized to find a summary age effect.

Results. This form of age-adjusted GH2000 score, a form of residual, has similar mean and variance as the original GH2000 score and, hence, the developed decision limits show negligible change when compared to the decision limits based on the original score. We also show that any further scale-transformation will not change the adjusted score any more. Hence the suggested adjustment is optimal for the given data. The summary age effect is homogeneous across the six scores, and so the generic adjustment of the GH2000 score formula is justified.

Conclusions. A final revised GH2000 score formula is provided which is independent of the age of the athlete under consideration.

10.7 Contributed - Communicating Statistics: Novel approaches to teaching statistical investigation

Thursday 8 September 9am – 10am

Is more statistics good for everyone?

Rhys Jones
Cardiff University

Cardiff Q-Step FE/Schools Initiative

There has been an overwhelmingly positive response to the FE/School engagement work, linked to developing and promoting context rich statistical courses, across England and Wales. These courses, primarily aimed at year 12 and 13 students, are focussed on the development of a new subject area called social analytics (the scientific investigation of social processes using statistical techniques and analysis). Individuals attending this sessions will gain practical insights into the innovative partnerships developed between universities, exam boards and schools/ FE colleges. An exemplification of the collaborative benefits will also be explored. The session will also focus on the pedagogical basis of the qualifications being created, the interdisciplinary nature and skills centred approach that has been adopted, and the educational impacts in terms of student attainment and achievement in other subject areas. The case will be made that developing students critical thinking and conceptual understanding of statistics, can have positive impacts on many other subject areas. These positive impacts include attitudes towards mathematics and statistics, as well as educational acheivement. Examples of context rich statistical worksheets will be disseminated to give the audience a greater understanding of the courses being developed.

10.7 Contributed - Communicating Statistics: Novel approaches to teaching statistical investigation

Thursday 8 September 9am – 10am

A Socio-Medical Data-based Introductory Statistics Course

Murray Aitkin
University of Melbourne

The course uses the database of 1296 families with a child in the book ``StatLab: an Empirical Introduction to Statistics'', by J.L. Hodges, D. Krech and R.S. Crutchfield (1975, McGraw-Hill), for a 24 class-hour course. However the theoretical approach to statistical analysis, for the University freshman class of social science/humanities students, is quite different. Students draw their own random samples from the data base (in the original version, by throwing two dice twice), and have to draw inferences about population proportions for binary variables (sex of baby, mothers' and fathers' smoking, baby birthweight -- low or not), and about relations between binary variables. These are expressed through plausible intervals (confidence or credible) derived from the binomial probabilities developed from the sampling procedure.

Students enjoy the practical aspects of the course, which were interspersed with set-piece lectures on the use and misuse of probability in the social, legal and medical world. The presenter enjoys giving the course and the good reactions from students. The direct data approach alerts students to the extensive data available, and also alerts them to the way the data were obtained, the need for representativeness, its achievement through random sampling, and the need and use for simple probability statements. Algebra is minimised, and all computations can be done quickly and simply, either without a computer or with a simple spread-sheet package.

The originality of the StatLab book data-based approach was handicapped by the highly traditional syllabus. The technical changes to the syllabus, and the inclusion in the course of real abuses of probability, make the course both more interesting and easier for students without a strong mathematical background. However the data-based approach is equally valuable for students with a strong mathematical background, with appropriate changes.

10.7 Contributed - Communicating Statistics: Novel approaches to teaching statistical investigation

Thursday 8 September 9am – 10am

Teaching Statistics to Train Statisticians: a practical based approach to developing statistical thinking and communication

Rosemary McNiece, Chris Wyman
Kingston University

This paper presents an approach to teaching undergraduate statistics which aims to train students in statistical thinking and problem solving using practical based examples and assessments. At undergraduate level, we routinely provide students with knowledge and understanding of theoretical and methodological concepts. However in training statisticians who are competent to enter the modern workplace we also have an obligation to impart practical skills in the applications of such concepts and a working knowledge of the many stages of conducting a statistical investigation from problem formulation through to communication of results. Hence while instruction in technical ability is important, instruction in the wider role of a statistician is equally if not more important. Students should be taught about the different stages in conducting a statistical investigation and make them aware of all the steps involved in producing reliable, accurate and comprehensible information as and when required.

However teaching these skills is not always easy in the typical undergraduate program where traditional resources and physical environments can limit options. We have developed a second year module which is taught from a very practical perspective drawing on blended learning techniques including peer led discussions, data analysis based problem solving and guided mini projects. We still use some more traditional lecture type sessions but these are kept to a minimum and group discursive learning is widely employed. Assessments are varied, using both individual and group exercises which build sets of skills leading to a final group project culminating in a semi-professional peer reviewed oral presentation.

Feedback from students has been very positive overall, although many admit that initially they found the interactive approach and peer led sessions to be daunting. The oral presentations provide a light hearted opportunity to interact with peers and critically review both their own and others students work.

11.1 INVITED - Methods & Theory: Probabilistic and statistical techniques for electrical power systems

Thursday 8 September 10.10am – 11.30am

Structural Price Modeling and Forward Curve Calibration in Modern Electricity Markets

Michael Coulon
University of Sussex

Rapidly-changing energy markets over recent years have led to important modeling challenges along with numerous opportunities to apply and extend techniques from statistics and mathematical finance. The literature in the area has grown significantly, with much interest in price or demand models, derivative pricing or hedging techniques, and optimal investment or operational decisions. We begin with a brief overview of the new and ongoing challenges of modern electricity markets, their unique features and how traditional quantitative finance approaches can be adapted. We present examples of suitable price modeling frameworks and applications, with a particular emphasis on the role of supply and demand variables and the use of structural models. Combining various approximation techniques, we derive closed-form formulas for forwards and options that capture the complex dependence structure between power, gas, coal and carbon emissions prices. In particular, based on work with industry partners, we propose a practical approach to the important problem of constructing and calibrating hourly forward curves in the German market in the face of a very diverse fuel mix, demand-side price elasticity, and the rapidly growing penetration of renewable generators.

11.1 INVITED - Methods & Theory: Probabilistic and statistical techniques for electrical power systems

Thursday 8 September 10.10am – 11.30am

Modelling and prediction of network time series

Marina Knight, Matthew Nunes, Guy Nason
University of York

Time series that arise on a graph or network are nowadays encountered in many scientific fields. In this work we propose a method for the modelling and prediction of such time series, with potentially complex characteristics. The proposed methodology is based on the lifting scheme, a technique with desirable properties, first proposed by Sweldens. By such a multiscale transformation, suitable for irregular data, the original network time series data is projected into a simpler, lower dimensional time series object which is easier to forecast. The technique is illustrated with a data set arising from an energy time series application.

11.2 INVITED - Medical Statistics: What do the experts believe? Use of prior elicitation to aid decision making in drug development and health technology assessment

Thursday 8 September 10.10am – 11.30am

Integrating multiple sources of evidence to estimate cost-effectiveness

Richard Lilford
University of Warwick

Cost-effectiveness is often estimated in health technology assessment by extrapolating from the results of randomised clinical trials. However, such evidence is often lacking for interventions which operate more 'distally' from the patient, such as service delivery interventions or public health policies. In these cases it is necessary to integrate multiple sources of evidence along a causal pathway to estimate the effects of interest. Specification of a Bayesian network enables identification of an estimator of these effects, which can be parameterised. Prior distributions for these parameters can be informed by both expert opinion and data, taking into account issues of internal and external validity of previous data sources. This is illustrated with two examples: electronic prescribing systems and seven day NHS.

11.2 INVITED - Medical Statistics: What do the experts believe? Use of prior elicitation to aid decision making in drug development and health technology assessment

Thursday 8 September 10.10am – 11.30am

Using prior elicitation to support decision making in drug development at GSK

Nicky Best, Nigel Dallow, Tim Montague
GSK

Since 2014, GlaxoSmithKline (GSK) has been using formal prior elicitation methods to support internal decision making and analysis in drug development. Prior elicitation is used to enable quantification of existing knowledge about an asset in the absence of directly relevant data. In this talk, I will provide a brief introduction to the methods that GSK have been using for prior elicitation, and discuss some of the benefits and challenges of embedding this process within a large pharmaceutical company. I will also gives some examples of how the elicited priors have been used at GSK, e.g. to quantitatively choose between competing clinical trial designs for the next stage of drug development, to explore staged development activities and to determine the merits of interim/futility assessments.

11.2 INVITED - Medical Statistics: What do the experts believe? Use of prior elicitation to aid decision making in drug development and health technology assessment

Thursday 8 September 10.10am – 11.30am

Bayesian prior elicitation: an application to the MYPAN trial in childhood polyarteritis nodosa

Lisa Hampson, John Whitehead, Despina Eleftheriou, Paul Brogan
Lancaster University

This presentation describes the process used to design the MYPAN trial, a randomised controlled trial for an inflammatory paediatric disease, childhood polyarteritis nodosa. The rarity of this condition means that a Europe-wide recruitment effort will likely yield a total sample size of 40 or fewer, even if patients are recruited over several years. For studies in such rare diseases, the sample size needed to meet a conventional frequentist power requirement is clearly infeasible. Rather, the expectation of any such trial has to be limited to the generation of an improved understanding of treatment options. The MYPAN trial, which is to compare an experimental treatment with control, classifying patient responses as success or failure, will follow a Bayesian design. We describe the process used to systematically elicit from clinicians their beliefs concerning treatment efficacy in order to establish Bayesian priors for unknown model parameters. We also outline how expert opinion on the relevance of results from a related historical trial was elicited and used to incorporate these data into priors based on opinion alone. As sample sizes are to be small it is possible to compute all possible posterior distributions that might result on termination of the MYPAN trial. Consideration of the extent to which opinion can be changed, even by the best feasible design, can help to determine whether a small trial is worthwhile and, if it is, the optimal randomisation ratio.

11.3 INVITED - Social Statistics: Statistical Analysis of Social Networks

Thursday 8 September 10.10am – 11.30am

Exploratory Data Analysis for Multiplex Networks with DISTATIS

Maria Prosperina Vitale, Giuseppe Giordano, Giancarlo Ragozini
University of Salerno

Multilayer network data arises when there exists more than one source of relationship for a group of actors (Kivelä *et al.*, 2014, and references therein). For such kind of data, the usual approach consists in dealing with multiple relations separately or in summing up the information embedded in all layers. This latter reduces the complexity of multiplex data and may lead to a loss of relevant information.

In the present contribution, aiming at visually explore the complex structure of multilayer networks, we propose to use factorial methods. These methods, in fact, have proven to be suitable to analyze the set of multiple relations seen as a whole complex structure (D'Esposito et al., 2014, Ragozini et al., 2015, Zhu et al. 2016). More specifically, given the data structure of one-mode multilayer networks, we propose to analyze the corresponding set of the adjacency matrices through the DISTATIS technique (Abdi *et al.*, 2012), which is an extension of the multidimensional scaling to a set of connected distance matrices. This technique, in a STATIS perspective (Lavit *et al.*, 1984), allows to represent the different kinds of relationships (inter-structures) in separate spaces and in a compromise space. By the use of DISTATIS we will be able to visually explore: *i)* the network structure in terms of actor similarity in each single layer, *ii)* the common structure of all layers, *iii)* the actor variations across layers, and *iv)* the similarities among layer structures. The proposed method will be discussed within some illustrative examples.

11.3 INVITED - Social Statistics: Statistical Analysis of Social Networks

Thursday 8 September 10.10am – 11.30am

Social network analysis

Nial Friel
University College Dublin

Exponential Random Graph models are an important tool in network analysis for describing complicated dependency structures. However, Bayesian parameter estimation for these models is extremely challenging, since evaluation of the posterior distribution typically involves the calculation of an intractable normalizing constant. This barrier motivates the consideration of tractable approximations to the likelihood function, such as pseudolikelihoods, which offer a principled approach to constructing such an approximation. Naive implementation of a posterior from a misspecified model is likely to give misleading inferences. We provide practical guidelines to calibrate in a quick and efficient manner samples coming from an approximated posterior and discuss the efficiency of this approach. The exposition of the methodology is accompanied by the analysis of real-world graphs. Comparisons against the Approximate Exchange algorithm of Caimo and Friel (2011) are provided, followed by concluding remarks.

11.3 INVITED - Social Statistics: Statistical Analysis of Social Networks

Thursday 8 September 10.10am – 11.30am

Efficient Bayesian computation for ERGMs

Alberto Caimo
Dublin Institute of Technology

Recent research in statistical social network analysis has demonstrated the advantages and effectiveness of Bayesian approaches to network data. In fact, Bayesian exponential random graph models (BERGMs) are becoming increasingly popular as techniques for modelling relational data in wide range of research areas. However, the applicability of these models in real-world settings is limited by computational complexity. In this talk, we review some of the most recent computational methods for estimating BERGMs as well as extended ERGM-based modelling frameworks for dynamic and heterogenous social networks.

11.3 INVITED - Social Statistics: Statistical Analysis of Social Networks

Thursday 8 September 10.10am – 11.30am

Imputation of missing social network data

Mark Huisman, Robert Krause, Tom Snijders
University of Groningen

We give an overview of the state of the art of imputation methods for social network data. Some simple procedures are introduced, as well as more elaborate methods based on exponential random graphs models and stochastic actor-oriented models. These latter models are the basis of multiple imputation methods, which will be examined in simulation studies.

11.4 INVITED - Data Science: Unpacking data visualisation

Thursday 8 September 10.10am – 11.30am

Data Visualisation and the Newsroom

Alan Smith
Financial Times

Data visualisations are increasingly being used by the media. This has brought with it investment and new hires at a time of contraction for much of the media industry. But what is the rationale for using data visualisation as a tool for mass communication? How do you build capability to deliver? And how does the pace of a newsroom affect the workflow for producing effective graphics. This talk will discuss some of the challenges, opportunities and broader trends of using data visualisation in a journalistic context. The author is Data Visualisation Editor at the Financial Times.

11.4 INVITED - Data Science: Unpacking data visualisation

Thursday 8 September 10.10am – 11.30am

Not Just Pretty Pictures: The Whats and Whys of Data Visualisation for Learning from Data

Jonathan Minton
University of Glasgow

Data Visualisations and Infographics are often thought of as synonyms, useful tools mainly for scientific public engagement, for puffing up, prettifying and promoting scientific findings to a wider audience. But not all infographics are data visualisations, and not all data visualisations are, or should aspire to be, infographics.

Starting with the grammar of graphics; definition of a data visualisation, and illustrated with examples of complex data visualisations of complex demographic data, this talk will argue for the role of data visualisations at all stages of scientific research, as a complement and corrective for occasional over-reliance on model-based inference, and as an essential component of iterative, quantitative research workflows that efficiently turn data into knowledge and insight. Data visualisations do not have to be pretty, but they do need to be effective.

11.5 INVITED - Environment Statistics: Papers from the Journal of the Royal Statistical Society

Thursday 8 September 10.10am – 11.30am

An SPDE Based Model for Large Space-Time Data Applied to Precipitation Postprocessing

Fabio Sigrist
Lucerne University of Applied Sciences and Arts

Increasingly larger spatial and spatio-temporal data sets are obtained, for instance, from remote sensing satellites or deterministic physical models such as numerical weather prediction (NWP) models.

We show that the solution of a stochastic advection–diffusion partial differential equation provides a flexible model class for spatiotemporal processes which is computationally feasible also for large data sets. The Gaussian process defined through the stochastic partial differential equation has, in general, a non-separable covariance structure. Its parameters can be physically interpreted as explicitly modelling phenomena such as transport and diffusion that occur in many natural processes in diverse fields ranging from environmental sciences to ecology. To obtain computationally efficient statistical algorithms, we use spectral methods to solve the stochastic partial differential equation. This has the advantage that approximation errors do not accumulate over time, and that in the spectral space the computational cost grows linearly with the dimension, the total computational cost of Bayesian or frequentist inference being dominated by the fast Fourier transform.

The incorporation of forecasts from an NWP model in a statistical model is generally called postprocessing. It provides calibrated and probabilistic forecasts. The proposed model is successfully applied to postprocessing of precipitation forecasts for northern Switzerland. In contrast with the raw forecasts from the numerical model, the post-processed forecasts are calibrated and quantify prediction uncertainty.

11.5 INVITED - Environment Statistics: Papers from the Journal of the Royal Statistical Society

Thursday 8 September 10.10am – 11.30am

Estimating the health benefit of reducing indoor air pollution in a randomized environmental intervention using principal stratification

Amber Hackstadt
Vanderbilt University Medical Center

Recent intervention studies targeted at reducing indoor air pollution have demonstrated both the ability to improve respiratory health outcomes and to reduce particulate matter (PM) concentrations in the home. However, these studies generally do not address whether it is the reduction in PM concentrations specifically that improves respiratory health. We propose a principal stratification (PS) approach to examine the extent to which an environmental intervention's effect on health outcomes coincides with its effect on indoor PM air pollution. The PS approach allows us to take advantage of the randomized experimental design and avoid bias in the environmental intervention effect estimates that may be introduced by an analysis that simply conditions on the post-treatment variable. We apply the method of principal stratification to data from a randomized air cleaner intervention designed to reduce indoor PM pollution in homes of children with asthma and allow the post-treatment environmental factor to be treated as discrete or continuous. We find that among children for whom the air cleaner would reduce indoor PM concentrations, the intervention would result in a meaningful improvement of asthma symptoms and estimate the effect of the intervention for this group to be larger than the overall effect estimate. This analysis demonstrates the usefulness of principal stratification for environmental intervention trials and its potential for much broader application in this area.

11.5 INVITED - Environment Statistics: Papers from the Journal of the Royal Statistical Society

Thursday 8 September 10.10am – 11.30am

Modelling heatwaves in central France: a case-study in extremal dependence

Hugo Winter, Jonathan Tawn, Simon Brown
EDF Energy

Heatwaves are phenomena that have large social and economic consequences. Understanding and estimating the frequency of such events are of great importance to climate scientists and decision makers. Heatwaves are a type of extreme event which are by definition rare and as such there are few data in the historical record to help planners. Extreme value theory is a general framework from which inference can be drawn from extreme events. When modelling heatwaves it is important to take into account the intensity and duration of events above a critical level as well as the interaction between both factors. Most previous methods assume that the duration distribution is independent of the critical level that is used to define a heatwave: a shortcoming that can lead to incorrect inferences. This talk will characterize a novel method for analysing the temporal dependence of heatwaves with reference to observed temperatures from Orleans in central France. This method enables estimation of the probabilities for heatwave events irrespectively of whether the duration distribution is independent of the critical level. The methods are demonstrated by estimating the probability of an event more severe than the 2003 European heatwave or an event that causes a specified increase in mortality. Additonal work concerning the effect of climate change on heatwaves will also be presented.

11.6 INVITED - Statistics in Sport: Recreational, Professional and Sports Betting

Thursday 8 September 10.10am – 11.30am

trackeR: Infrastructure for Running and Cycling Data in R

Hannah Frick, Ioannis Kosmidis
University College London

The use of GPS-enabled tracking devices and heart rate monitors is becoming increasingly common in sports and fitness activities. The trackeR package aims to fill the gap between the routine collection of data from such devices and their analyses in a modern statistical environment like R. The package provides methods to read tracking data and store them in session-based, unit-aware, and operation-aware objects of class "trackeRdata". The package also implements core infrastructure for relevant summaries and visualisations, as well as support for handling units of measurement. There are also methods for relevant analytic tools such as time spent in zones, work capacity above critical power (known as W'), and distribution and concentration profiles. A case study illustrates how the latter can be used to summarise the information from training sessions and use it in more advanced statistical analyses.

11.6 INVITED - Statistics in Sport: Recreational, Professional and Sports Betting

Thursday 8 September 10.10am – 11.30am

A statistical model of the effect of training on performance in road cycling

Phil Scarf
University of Salford

Power output and heart rate data are routinely collected by riders during training, testing and competition. This paper considers whether such data are useful for quantifying the response of performance to training, and ultimately whether a performance-training model can be used to plan training. A classic performance-training model (the Banister model) is estimated for ten male, competitive cyclists. The model and its estimation are described. There are a number of issues with the model. The first is whether the model parameters are sufficiently well estimated for practical use. The second and more fundamental is that the notion of training capacity is absent from the Banister model. A potential solution based on the critical power model (a model that specifies the maximum power output that can be sustained for a particular duration) will be presented.

11.7 INVITED - Communicating Statistics: Communication statistics to non-specialist university students

Thursday 8 September 10.10am – 11.30am

Livening up statistics teaching rooms by “LOL"

Meena Kotecha
The London School of Economics and Political Science

This interactive presentation will commence with demystifying “LOL”. It will share the key features and the impact of an innovative instruction method that was developed during a longitudinal study conducted specially to understand and address the following challenge:

Non-specialists (undergraduates enrolled on a range of degree programmes other than statistics) generally display negative attitudes towards the mandatory statistics courses which they are required to study as core modules of their respective degree programmes. Moreover, they report negative emotions linked to “statistics anxiety” despite having successfully completed A-Level mathematics or equivalent. This can impede their enthusiasm to engage with the courses, adversely affecting their academic performance and employment profile as a result. This presents a challenge to academics involved in delivering such courses.

The presentation will touch upon the research methodology specially developed for this study. Furthermore, the delegates will be able to hear from a few non-specialists about how the proposed instruction method transformed their attitudes towards engaging with statistics.

Academics from all related disciplines should be able to apply the proposed techniques to delivering any quantitative courses designed for non-specialists. Furthermore, this should be of interest to statistics education researchers and all interested in the theme.

11.7 INVITED - Communicating Statistics: Communication statistics to non-specialist university students

Thursday 8 September 10.10am – 11.30am

University students reading life sciences and statistical thinking

Matina Rassias
University College London (UCL)

In an era of continuous changes and avalanche of information, education ought to lead the way in dealing with the challenges. It is the educators' responsibility to have a profound impact in shaping the forward thinking not only of the specialist but most importantly of the non-specialist students in a discipline.

The focus of this presentation is to address some of the notable challenges related to learning and teaching Statistics to non-specialists and to explore ways to introduce and gradually develop a statistical oriented mind-set, considering in particular Life-Sciences students. Recognising the undeniable benefits of a research-based education, how can educators facilitate the development of an inclusive learning community where students learn through dialogue, and active critical enquiry? How can educational technologies improve the experiences of both educators and students and assist towards connections beyond the narrow classroom borders? These will be some of the questions we aim to address and share our experience with the attendees.

11.7 INVITED - Communicating Statistics: Communication statistics to non-specialist university students

Thursday 8 September 10.10am – 11.30am

Stats buddy: helping non-statisticians to help their students learn medical statistics

Jamie Sergeant
University of Manchester

University students in the medical and human sciences often have to learn statistics as part of their course, for example to enable them to understand or perform the analysis of data or to enable them to design and conduct research projects. However, this teaching and learning may be delivered by non-statisticians who may not have the expertise, experience or confidence of subject specialists.

This presentation will describe two sources of support for early career academics at the University of Manchester who are not statisticians but who deliver statistics teaching and learning to students in the medical and human sciences. The first is a training session to enable the sharing of experiences and the exploration of different approaches to making statistics engaging for students. The second is a "buddying" scheme in which the non-specialist educators can be paired with statistician buddies who can help them to meet the needs of their students.

12.1 Contributed - Methods & Theory: Predictive Inference

Thursday 8 September 12noon – 1pm

Nonparametric predictive inference for future order statistics

Hana Alqifari, Frank Coolen
Durham University

Nonparametric predictive inference (NPI) is a powerful frequentist statistical framework, with inference explicitly in terms of future observations. NPI is based on the assumption $A_{(n)}$, proposed by Hill in 1968, which provides a partially specified predictive probability distribution for a single future observation given past data. Most of the theory on NPI has been so far restricted to a single future observation, although multiple future observations have been consider for some NPI methods for statistical process control. In this research further theory has been developed on NPI for multiple future order statistics. We focus on the order statistics of m future observations given n past data. An important feature of NPI is that the interdependence of the future observations is explicitly taken into account. NPI is suitable for prediction in the situation if there is hardly any knowledge about the underlying distribution of the observations. We derive precise probabilities for some events involving one or more future observations. The assumption of $A_{(n)}$ is not sufficient to derive precise probabilities for many events of interest, but it provides optimal bounds for probabilities for all events of interest involving future observations. These bounds are lower and upper probabilities in the theory of imprecise probability.

We present NPI for a specific order statistic of m future real-valued observations as well as joint and conditional probabilities for the future order statistics. We show how pairwise comparison of different groups can be based on such future order statistics, and we briefly discuss some generalizations including multiple comparisons.

12.1 Contributed - Methods & Theory: Predictive Inference

Thursday 8 September 12noon – 1pm

Nonparametric predictive inference for diagnostic test thresholds

Manal Alabdulhadi, Frank Coolen, Tahani Coolen-Maturi
Durham University

The accuracy of diagnostic tests relates to the ability of the tests to distinguish between diseased and healthy individuals. Providing good methods for defining the accuracy of diagnostic tests assist physicians to detect the probability of disease for their patients. In 2-Group and 3-Group ROC analysis, setting thresholds for classification is often the most important decision. The standard uses the maximisation of the Youden index, a global measurement of diagnostic accuracy that is the vertical distance between ROC curve value corresponding the threshold *c* and the point on the diagonal line. In this work, we consider an alternative to the maximisation of the Youden index, by explicitly considering the use of the classification procedure for a specific number of future patients. We consider nonparametric predictive inference (NPI), which is a powerful statistical framework that yields direct probabilities for one or m future observations, based on *n* observations for related random quantities. It provides bounds for probabilities for events of interest; these are lower and upper probabilities in the theory of imprecise probability. In this work, we introduce 2-Group and 3-Group predictive method to select optimal diagnostic thresholds in order to have the best classification of one or more future persons on the basis of their test results. We find that the optimal thresholds sometimes lead to other value of thresholds with different number of m future patients. We generalize the Youden index by applying our method to the Youden index and maximising the sum of the probabilities of correct classification for the different groups. Comparison between our method and generalization of the Youden index is discussed.

12.1 Contributed - Methods & Theory: Predictive Inference

Thursday 8 September 12noon – 1pm

Predictive Inference with Copulas for Bivariate Data

Tahani Coolen-Maturi, Frank Coolen, Noryanti Muhammad
Durham University

Nonparametric predictive inference (NPI) is a statistical approach with strong frequentist properties, with inferences explicitly in terms of one or more future observations. NPI is based on relatively few modelling assumptions, enabled by the use of lower and upper probabilities to quantify uncertainty. While NPI has been developed for a range of data types, and for a variety of applications, thus far it has not been developed for multivariate data. This paper presents the first study in this direction. Restricting attention to bivariate data, a novel approach is presented which combines NPI for the marginals with copulas for representing the dependence between the two variables. As an example application of our new method, we consider accuracy of diagnostic tests with bivariate outcomes, where the weighted combination of both variables can lead to better diagnostic results than the use of either of the variables alone. As this is the first research into developing NPI-based methods for multivariate data, there are many related research opportunities and challenges, which we briefly discuss.

12.2 Contributed - Medical Statistics: Analysis of Clustered Data

Thursday 8 September 12noon – 1pm

Bayesian Hierarchical Modelling of the Intra-cluster Correlation Coefficient for a Cluster-Randomised Trial

Svetlana Tishkovskaya, Chris Sutton, Lois Thomas, Michael Leathley, Caroline Watkins
University of Central Lancashire

Objectives

A major difficulty in planning a cluster-randomised trial (CRT) is obtaining a robust estimate of the intra-cluster correlation coefficient (ICC). The aim of the study was to use Bayesian hierarchical modelling to construct a distribution of the ICC to inform the design of a CRT.

Methods and Models

We adopted the method of combining ICC values in the Bayesian framework suggested by Turner, Thompson and Spiegelhalter (2005), using Bayesian hierarchical modelling to construct a distribution of the ICC to aid planning of a CRT of incontinence care following stroke.

Through a literature search, studies reporting ICC estimates with varying degrees of relevance to the planned trial and to its primary outcome (frequency of urinary incontinence) were identified. To allow for uncertainty in ICC estimates, the Bayesian hierarchical model requires separate weights for each study and for each outcome. We performed an exercise by which a team of expert reviewers with a range of relevant expertise assigned weights for each trial and outcome. This knowledge elicitation process produced a set of weights which were incorporated into the model to construct a distribution of a targeted ICC combining information from the individual ICC estimates.

Results and Conclusions

The ICC distribution constructed in a Bayesian framework was based on multiple ICC estimates of varying relevance (22 estimates from 8 studies). This approach improved the robustness of the estimation of the ICC for the primary outcome, and hence the sample size, for a planned CRT by quantifying uncertainty about the ICC and incorporating expert knowledge into the model.

12.2 Contributed - Medical Statistics: Analysis of Clustered Data

Thursday 8 September 12noon – 1pm

Risk prediction for clustered data with few events

Menelaos Pavlou, Gareth Ambler, Shaun Seaman, Rumana Omar
University College London

Risk prediction models typically use a number of predictors based on patient characteristics to predict health outcomes. When there are few events compared to the number of predictor variables, model overfitting can be a serious problem resulting in poorly calibrated predicted risks. In such cases, penalised regression methods such as ridge and lasso have been shown to produce models with superior predictive performance than the standard regression model fitted by maximum likelihood (ML). In this work we focus on risk prediction models for clustered data (for example patients clustered within hospitals) with few events. For binary outcomes, a logistic random intercepts regression model which accounts for the intracluster correlation (ICC) is commonly used to model the association between predictors and the outcome, and can provide both cluster-specific and marginal predictions. However, just as in the case of independent data, the standard random intercepts model (fitted using ML) is susceptible to model overfitting. Bayesian analogues of ridge and lasso, which impose shrinkage priors on the regression coefficients, can be easily adapted to the clustered-data setting by including random intercept terms. In simulation studies based on real data, with varying degrees of ICC and cluster sizes, we evaluated the predictive performance of Bayesian lasso with random intercepts and found that it produces better calibrated models with lower mean square predictive error than those obtained by the standard random intercepts model. In the presence of noise predictors, Bayesian Lasso can simultaneously perform variable selection and shrinkage, and was found be a better alternative than stepwise methods (e.g. backwards elimination). However, when ICC is relatively small (<0.15), standard ridge or lasso which ignore clustering can produce models with predictive performance similar to that of Bayesian lasso with random intercepts.

12.2 Contributed - Medical Statistics: Analysis of Clustered Data

Thursday 8 September 12noon – 1pm

The role of cluster size and intra-cluster correlations when adjusting for covariates in the analysis of cluster randomised trials

Neil Wright
Queen Mary University of London

Reports of clinical trials often include adjusted analyses, which incorporate covariate data into the analysis model. Adjusting for covariates can increase the power to detect a treatment effect, without the need to increase sample size. In individually randomised trials the main reason to adjust for a particular covariate is that it is expected to be strongly associated with the primary outcome. The larger the association between covariate and outcome, the greater the increase in power achieved from an adjusted analysis.

In a cluster randomised trial (CRT) pre-existing groups, called clusters, of participants are randomised to treatment arms. A valid analysis of a CRT must take into account the clustered structure of the data, for example by using a mixed effects model.
Choosing covariates for an adjusted analysis of a CRT is more complicated because covariates exist at both the cluster level and individual level, and adjustment for an individual level covariate can affect the residual variance of the outcome at both levels.

Using results from simulations, and analytic investigation, I show how cluster size and the intra-cluster correlations of the covariate and outcome affect power and precision in adjusted analyses of CRTs using linear mixed effects models. I also consider logistic mixed effects models and show how adjusting for individual level or cluster level covariates affect the treatment effect that is to be estimated and the precision of that estimate.

12.3 INVITED - Research Students Conference Prize winners

Thursday 8 September 12noon – 1pm

Clustering Nonstationary Circadian Plant Rhythms using Locally Stationary Wavelet Representations

Jessica Hargreaves, Marina Knight, Jon Pitchford, Seth Davis
University of York

How does soil pollution affect the plant's circadian clock? Are there any differences between how the clock reacts when exposed to different elements of the periodic table? If so, can we characterise these differences?

We approach these questions by analysing and modelling circadian plant data, where the levels of expression of a luciferase reporter gene were measured at regular intervals over a number of days after exposure to different concentrations of lithium.

A key aspect of circadian data analysis is to determine whether a time series (derived from experimental data) is "rhythmic" and, if so, to determine the underlying period using Fourier analysis. However, we show that this data is nonstationary. Therefore, we assert that the current methods to analyse such data are inadequate and propose a new framework to deal with the challenges of circadian data.

In this talk, we propose a method for clustering nonstationary time series using a bias-corrected nondecimated wavelet transform. Wavelets are chosen as they are ideally suited to identify discriminant local time and scale features. We model the observed plant luciferase signals as realisations of locally stationary wavelet processes. This allows us to define and rigorously estimate the plant-specific individual evolutionary wavelet spectra, hence yielding a reliable time-scale decomposition for each process variance. The estimated spectra are used to cluster the data into homogenous groups. However, as such data are high dimensional, we propose an intermediate dimension reduction step. This is achieved by treating the estimated spectra as “images” and performing a functional principal components analysis. An effective dissimilarity measure is then used for the projected spectra and we show that classical clustering algorithms, using this measure, effectively differentiate among rhythmic behaviour. Finally, we demonstrate the advantages of our methodology on the circadian data.

12.3 INVITED - Research Students Conference Prize winners

Thursday 8 September 12noon – 1pm

An Alternative Approach to Calibration in Survey Sampling

Gareth Davies, Jonathan Gillard, Anatoly Zhigljavsky
Cardiff University

Survey calibration is one of the key methodologies used by many international statistical offices. For example, the Office for National Statistics uses calibration in the production of quarterly estimates of the UK unemployment rate. Many of the other statistics that appear in the news headlines are also obtained using survey calibration.

Initially, a 'design' weight is assigned to each respondent in the sample. Whilst the design weights may be sufficient in certain estimation problems, many problems in official statistics require weights that give consistent estimates with known totals from other surveys or the Census. Survey calibration is used to adjust the design weights.

The classical calibration procedure seeks to minimize the deviation between the new, calibrated weights and the initial, design weights. The calibrated weights are then used to form population estimates. There are many so-called ‘distance functions’ that can be used to assess the deviation between the design weights and the calibrated weights, and we shall discuss the properties of several of these functions during the talk.

However, in practice, the variance of the calibration estimator is of more interest than the weights themselves. We propose an alternative approach to calibration that minimizes the mean squared error of the calibration estimator. From a practical perspective, this approach has many advantages over standard calibration techniques. We shall illustrate this for several examples, and show that the proposed method leads to more consistent estimates.

12.3 INVITED - Research Students Conference Prize winners

Thursday 8 September 12noon – 1pm

A Markov Random Fields approach to the Gating of Flow Cytometry Data

Kevin Brosnan, Kevin Hayes, Norma Bargary
University of Limerick

Introduction

Flow cytometry is a technology that simultaneously measures and analyses multiple physical and chemical characteristics of single cells as they flow in a stream through a beam of laser light. This technology has become an emerging state-of-the-art device in microbiology and dairy science, and is also used extensively in medical diagnostics. The gating stage of analysis, the identification of homogeneous cell populations, is performed using expert opinion rather than by employing a unified statistical framework. The increased volume and complexity of flow cytometry data resulting from advances in the technology greatly boosts the demand for reliable statistical methods and accompanying software implementations for analysis. The objective of this research is to provide a statistically robust methodology for the gating of flow cytometry data which moves beyond the expert-driven approach currently employed.

Methodology

One aspect of flow cytometry data which has not been properly exploited is the integer valued nature of such data, resulting from the technology used to record it. Given this structural layer of flow cytometry data each pair of recorded variables can be visualised as a 2-dimensional image. Markov Random Fields (MRF) have been used extensively for image modelling and in recent years for drawing inferences from images. Our approach is to utilise this structure by applying an MRF approach to allow the constructed images to be segmented into regions, each containing a unique sub-population of the recorded cells. MRF also provide a probabilistic foundation to the grid of flow cytometry data, allowing additional inferences to be drawn from the image data.

Results

The methodology is applied to the rituximab data which appears frequently throughout the flow cytometry literature for the purpose of demonstrating statistical methods. The results provide a visual and statistical comparison of the sub-populations identified utilising the current expert-driven approach and the proposed statistical approach.

12.4 Contributed - Data Science: Model selection

Thursday 8 September 12noon – 1pm

Model selection in sparse multi-dimensional contingency tables

Susana Conde, Gilbert MacKenzie
University of Manchester

We compare some model selection methods in sparse high-dimensional contingency tables. These include the LASSO, a new Smooth LASSO which can be calculated in a standard regression analysis mould and a classical backwards elimination algorithm, as used in standard software packages. We undertake a simulation study, programmed in the R software package, which explores the ability of the three algorithms to identify the correct model as the sample size and the dimension of the table increases. Finally, we analyse an extensive set of comorbidity data arising in a study of obesity. The results show that the LASSO has difficulty finding the true model. Backwards elimination performs best for small sample sizes while the Smooth LASSO performs better as the sample size increases. Overall, the findings do not support the use of the standard LASSO as a mainstream method of model selection in sparse high-dimensional contingency tables. It is well known that the LASSO lacks the oracle property and our results tend to confirm this finding as it persistently over-fits effects. However, our results also suggest that the Smooth LASSO may possess the oracle property - a rather surprising finding which requires further investigation.

12.4 Contributed - Data Science: Model selection

Thursday 8 September 12noon – 1pm

Markov chain sampling for algorithmic leveraging

Keith Knight
University of Toronto

Sampling is often used for approximating least squares estimation with large-scale data sets. One popular method, called algorithmic leveraging, uses the leverage scores (the diagonals of the "hat" matrix) as an importance sampling distribution. While this method allows influential observations - those with high leverage scores - to be well-represented in the sample as a whole, their distribution within a sample may lead to bias in the resulting least squares estimates. Using the representation of the least squares estimate as a weighted sum of elemental estimates as a motivation, we propose a refinement of algorithmic leveraging using a Markov chain, which employs both the leverage and cross-leverage scores (the off-diagonal elements of the "hat" matrix) to assure a more even distribution of high leverage observations within a sample.

12.4 Contributed - Data Science: Model selection

Thursday 8 September 12noon – 1pm

Iteratively reweighted adaptive lasso for conditional heteroscedastic time series with applications to AR–ARCH type processes

Florian Ziel
European University Viadrina Frankfurt (Oder)

Shrinkage algorithms are of great importance in almost every area of statistics due to the increasing impact of big data. Especially time series analysis benefits from efficient and rapid estimation techniques such as the lasso. However, currently lasso type estimators for autoregressive time series models still focus on models with homoscedastic residuals. Therefore, an iteratively reweighted adaptive lasso algorithm for the estimation of time series models under conditional heteroscedasticity is presented in a high-dimensional setting. The asymptotic behaviour of the resulting estimator is analysed. It is found that the proposed estimation procedure performs substantially better than its homoscedastic counterpart. A special case of the algorithm is suitable to compute the estimated multivariate AR–ARCH type models efficiently. Extensions to the model like periodic AR–ARCH, threshold AR–ARCH or ARMA–GARCH are discussed. Finally, different simulation results and applications to electricity market data and returns of metal prices are shown.

Ref: Ziel, F. (2015). Iteratively reweighted adaptive lasso for conditional heteroscedastic time series with applications to AR–ARCH type processes. *Computational Statistics & Data Analysis*.

12.5 Contributed - Environment Statistics: Cattle / Air Pollution

Thursday 8 September 12noon – 1pm

A common singular spectrum analysis of bovine tuberculosis incidence in Great Britain

Theo Pepler, Rowland Kao
University of Glasgow

Singular spectrum analysis is used to decompose a univariate time series into orthogonal components. Analogous to principal components, the estimated singular spectra can be interpreted as accounting for observed features in the time series data. If the time series has periodic oscillations and/or longer term trends, these signals are usually accounted for by the first small number of spectra. The time series can be reconstructed in the subspace spanned by a selected number of spectra, to filter out noise in the observed data.

Under a common principal components model, the singular spectra of several time series can be compared to determine whether the different series have common oscillatory features. This type of analysis is known as common singular spectrum analysis (CSSA).

In this presentation, an application of CSSA to data on bovine tuberculosis (bTB) incidence in cattle herds in different regions of Great Britain is presented. Common characteristics (as described by estimated common singular spectra) in the different time series are noted and interpreted in the context of bTB epidemiology. Interesting similarities and differences between singular spectra of high and low bTB-risk areas in Great Britain are discussed. While bTB incidence in England and Wales seem to have some sources of variation in common, it appears that bTB incidence in Scotland can be attributed largely to noise components.

12.5 Contributed - Environment Statistics: Cattle / Air Pollution

Thursday 8 September 12noon – 1pm

Modelling non-linear exposure-response and non-linear lagged-response with respect to exposure to air pollution and cause-specific mortality risk

Matthew Gittins, Roseanne McNamee, Raymond Agius
University of Manchester

Background: - Previous attempts to model delayed (lagged) effects of air pollution and mortality risk have assumed a linear exposure-response across the entire lag period. Extrapolating this linearity assumption across a lag period of 20+ days based on previous exposure-response models of same day or previous day exposure, may cause misrepresentation of the true relationship particularly under differing causes of death. Here we investigate the assumption of linearity in causes-specific mortality and identify the most appropriate model to represent the data.

Methods: - Individual cause-specific mortality data under a time-stratified case-crossover design with matched exposure data is modelled using a conditional logistic regression. Natural cubic splines flexibly represent the exposure data and the lag period separately. The two sets of terms representing the natural cubic splines are combined into non-linear distributed lag model and included as terms in the regression model. Surface plots for each cause of death represent the change in risk across both exposure and lag and allow for visual inspection of the two relationships simultaneously.

Results: - 3D surface plots of temperature and pollution respectively, both largely confirmed the non-linear and linear exposure-response assumption across the entire 30 day lag period. Risk across the lag period due to pollution exposure indicated delayed peaks that may be more suitably modelled using cubic or greater polynomials.

Conclusions:- Pollution exposure-response can largely be assumed to be linear across the entire lag period, confirming previous results and allowing for simpler modelling techniques to represent lag-response effects making interpretation easier.

12.5 Contributed - Environment Statistics: Cattle / Air Pollution

Thursday 8 September 12noon – 1pm

A compositional mixed model for methane production from cattle

Javier Palarea-Albaladejo, John A Cooke, Ian M
Biomathematics & Statistics Scotland

Background

Data of compositional nature frequently arise in natural and biological sciences (e.g. chemical compositions, species abundance profiles or behavioural patterns). They are intrinsically co-dependent positive amounts carrying only relative information, and so they are typically expressed as percentages of a whole or equivalent units. These particularities, when ignored, have been shown to cause both technical and interpretability issues in data analysis such as singularity and multicollinearity in linear models, results dependent on the scale and size of the composition and spurious correlations.

Methods

Mapping compositions onto the real space by isometric log-ratio coordinates has been proved to be a suitable approach. However, there are infinitely many ways to define these and so exploratory analysis and expert knowledge are essential when interpretable coordinates are sought. Besides, results should not depend on the mapping chosen. A compositional linear mixed model (CoDA LMM) is introduced aimed at assessing the role of the volatile fatty acid (VFA) composition resulting from the fermentation of feed in ruminant forestomach in the methane production of cattle. The data come from a number of field experiments, and include factors such as diet type, dry matter intake and metabolisable energy.

Results

We show how a CoDA LMM can be formulated in this context. Estimations associated with the random effects, as well as other performance measures, are shown to be invariant under changes in the log-ratio mapping. Biologically, this type of modelling is interesting because methane produced by cattle contributes substantially to the greenhouse gases implicated in global warning. We demonstrate interesting relationships between VFAs and identify the balance between Butyrate and Propionate as a primary driver of methane emissions, once the effect of the other variables has been accounted for. Namely, the higher the relative weight of Butyrate in relation to Propionate the higher the predicted methane production.

12.6 Contributed - Sport Statistics: Modelling Championship Performance

Thursday 8 September 12noon – 1pm

Modelling perfectly competitive football leagues

John Fry
Sheffield Hallam University

In the light of enhanced TV deals and a seemingly increasingly unpredictable English Premier League season we model a perfectly competitive football league. Our admittedly simple model is able to re-produce the much-vaunted 40-point threshold as a target for avoiding relegation. Points targets are also derived for Champions League qualification and League Championship title winners. Applications to testing the efficiency of football labour markets are also discussed.

12.6 Contributed - Sport Statistics: Modelling Championship Performance

Thursday 8 September 12noon – 1pm

Formula for success: Multilevel modelling of Formula One Driver and Constructor performance, 1950-2014

Andrew Bell, James Smith, Clive Sabel, Kelvyn Jones
University of Sheffield

This paper uses random-coefficient models and (a) finds rankings of who are the best formula 1 (F1) drivers of all time, conditional on team performance; (b) quantifies how much teams and drivers matter; and (c) quantifies how team and driver effects vary over time and under different racing conditions. The points scored by drivers in a race (standardised across seasons and Normalised) is used as the response variable in a cross-classified multilevel model that partitions variance into team, team-year and driver levels. These effects are then allowed to vary by year, track type and weather conditions using complex variance functions. Juan Manuel Fangio is found to be the greatest driver of all time. Team effects are shown to be more important than driver effects (and increasingly so over time), although their importance may be reduced in wet weather and on street tracks. A sensitivity analysis was undertaken with various forms of the dependent variable; this did not lead to substantively different conclusions. We argue that the approach can be applied more widely across the social sciences, to examine individual and team performance under changing conditions.

12.6 Contributed - Sport Statistics: Modelling Championship Performance

Thursday 8 September 12noon – 1pm

Consistent Bradley-Terry model for fair standings in football leagues and other round-robin tournaments

David Firth, Heather Turner
University of Warwick

A generalization of the Bradley-Terry model is developed which yields end-of-season agreement with league standings, in balanced round-robin leagues with simple win-draw-loss points systems such as the 3-1-0 system used in association football. This allows mid-season "strength of schedule" differences to be eliminated coherently, in order to produce match-by-match league standings that are potentially more meaningful than the familiar ranking based on accumulated points. Results from several seasons of major European football leagues are used to assess the appropriateness of the model, and to calibrate it in aspects such as the relative frequency of draws and the "home advantage" effect. Effective presentation of mid-season league standings is achieved through expected end-of-season league points; a Dirichlet-prior shrinkage penalty, again calibrated from past European football league seasons, helps to keep such expected points totals realistic.

13.1 INVITED - Methods & Theory: Bench-marking in Clustering

Thursday 8 September 2pm – 3.20pm

Benchmarking in cluster analysis: An introduction to the International Federation of Classification Societies (IFCS) Cluster Benchmark Data Repository

Nema Dean
University of Glasgow

The IFCS Task Force on Benchmarking -

The research area of cluster analysis is constantly expanding with proposals of novel methods. However, to justify their existence, new methods are expected to be evaluated in a number of ways. Usually the paper introducing a new method will include a demonstration of application of the method to artificial and real world datasets, ideally in conjunction with a comparison of the method's performance with a selection of competing methods. One of the most popular sources of data sets used in these types of paper is the UCI machine learning repository. However, the data sets from this type of repository are usually classification data, which, while useful for benchmarking a classification method, may not be suitable for evaluating clustering methods. For example, there may be other feasible groupings than the classification provided. In order to address the lack of a data repository dedicated to providing data for benchmarking clustering methods, a new International Federation of Classification Societies (IFCS) repository is being introduced which provides tools for collection and distribution of such data sets. As well as real data, additional tools for generation of artificial datasets with desirable properties are provided. Part of what distinguishes this repository from others is the detailed questionnaire that accompanies each data upload that allows for high quality metadata relevant to clustering purposes to be provided with each data set. Such metadata includes the standard background information and previous usages, as well as relationships between variables, known group structures, the goal behind the clustering, etc. The ultimate goal is to have a collection of data sets that are routinely used for cluster algorithm comparison such that results are reproducible and comparable across scientific publications by different authors allowing for proper cumulative building of advances in clustering research.

13.1 INVITED - Methods & Theory: Bench-marking in Clustering

Thursday 8 September 2pm – 3.20pm

Benchmarking in cluster analysis: Preview of a white paper

Iven Van Mechelen
University of Leuven (Belgium)

To achieve scientific progress in terms of building a cumulative body of knowledge, careful attention to benchmarking is of the utmost importance. This means that proposals of new methods of data pre-processing, new data-analytic techniques, and new methods of output post-processing, should be extensively and carefully compared with existing alternatives. To date, benchmarking and recommendations for benchmarking have been frequently seen in the context of supervised learning. Yet, unfortunately, there has been a dearth of guidelines for benchmarking in an unsupervised setting, with the area of clustering as an important subdomain. To address this problem, a Task Force within the International Federation of Classification Societies (IFCS) is currently preparing a white paper on benchmarking in cluster analysis. In this white paper discussion is given to general and fundamental conceptual underpinnings of benchmarking in the field of cluster analysis by means of simulated as well as empirical data. Subsequently, the specifics of how to address benchmarking questions in clustering are dealt with, and foundational recommendations are made. In the present talk I will offer a sneak preview of this white paper.

13.3 INVITED - Social Statistics: Paradigms in non-random sampling

Thursday 8 September 2pm – 3.20pm

Missing Non-Voters and Misweighted Samples: Explaining the 2015 Great British Polling Miss

Jonathan Mellon, Christopher Prosser
University of Oxford

The pre-election polls for the 2015 UK General Election missed the final result by a considerable margin: underestimating the Conservative Party and overestimating Labour. We analyse evidence for five theories of why the polls missed using British Election Study data. We find limited evidence for systematic vote intention misreporting, late swing, systematically different preferences among “don’t knows” or differential turnout of parties’ supporters. By comparing the BES face-to-face probability sample and online panel, we show that the online survey’s polling error is primarily caused by under-sampling non-voters, then weighting respondents to represent the general population. Consequently, demographic groups with a low probability of voting are over-weighted within the voter subsample. Finally, we show that this mechanism is likely partially responsible for the over-estimate of the Liberal Democrats in 2010, showing that this is a longstanding problem with British polls.

13.3 INVITED - Social Statistics: Paradigms in non-random sampling

Thursday 8 September 2pm – 3.20pm

Who Tweets? First Steps in Understanding and Evaluating Representation on Twitter

Luke Sloan
Cardiff University

Twitter provides social scientists with a rich vein of data on attitudes, reactions, sentiment and networks at a scale we are only just learning to deal with – but whilst the amount of data creates new opportunities, the paucity of information about *who tweets* prevents us from exploiting them. In this talk we outline recently developed methods for estimating users demographic proxies in an attempt to understand the Twitter population thus laying the foundations for further investigation around representation and bias. We discuss the details of automating demographic categorisation using computational methods alongside the strengths and weaknesses of the approach. Methodological limitations aside, we tentatively conclude that Twitter users are not proportionally representative of the UK population.

13.4 INVITED - Data Science: Applications of machine learning in biostatistics

Thursday 8 September 2pm – 3.20pm

Estimating the comparative effectiveness of longitudinal treatment regimes: an application of targeted maximum likelihood estimation in critical care

Noemi Kreif, Linh Tran, Richard Grieve, Bianca DeStavola, Robert Tasker, Maya Petersen
London School of Hygiene and Tropical Medicine

Routinely collected longitudinal data sources offer new opportunities for the evaluation of sequential treatment interventions. Here, treatment decisions are updated according to new information available on the prognosis of the individual patient, necessitating the adjustment for time-depending confounding. Longitudinal targeted maximum likelihood based estimation (TMLE) is a double-robust method that can be coupled with machine learning, in particular the Super Learner, a cross-validation based estimator selection approach. To date there has been little work on evaluating the implications of applying this method in practice. The aim of this paper is to provide a critical examination of longitudinal TMLE, and contrast it with inverse probability of treatment weighting, while investigating the causal effect of nutritional interventions on clinical outcomes of critically ill children. We estimate the risk of being discharged alive from the pediatric intensive care unit by a given day, under a range of static and dynamic regimes. We find that before adjustment, patients who follow the static regime “Never feed", are discharged earlier than patients who follow the regime “Feed from day 3”, or “Feed when off mechanical ventilation”. After adjusting for baseline and time varying confounders, most of this difference disappears. We conclude that TMLE offers a flexible estimation approach that merits wider application in comparative effectiveness research.

13.4 INVITED - Data Science: Applications of machine learning in biostatistics

Thursday 8 September 2pm – 3.20pm

Applying Machine Learning Methods to Improve Analyses of Medical Research Studies

Daniel Stahl
King's College London

The aim of this study is to assess the usefulness of machine learning algorithms for applications in medical research as an alternative to classical statistical inference methods. The usefulness of a regularized regression and relatively simple regularized discriminant function analyses integrated with feature variable selection based on t-scores will be assessed by reanalyzing an event-related brain potential (ERP) dataset from infants at high or low risk of developing autism. Results will be compared to results from support vector machines. Recording Event-related brain potentials is a non-invasive method of measuring brain activity during cognitive processing with high temporal resolution. The standard analysis of averaged ERP measurements usually involves a large number of univariate mean group comparisons resulting in a multiple testing problem. Machine learning methods combined with cross-validation methods allow assessing the predictive performance of a derived model, thereby avoiding multiple testing problems. The analyses showed that both machine learning methods successfully discriminated above chance between groups of infants at high and low risk of a later diagnosis of autism and correlation-adjusted t-scores allowed identifying key variables, which separated the two groups.

13.4 INVITED - Data Science: Applications of machine learning in biostatistics

Thursday 8 September 2pm – 3.20pm

Targeted Statistical Learning for Health Care Spending

Sherri Rose
Harvard

Adjusting for health conditions is ubiquitous in health care. The federal government, as well as health plans and provider organizations, routinely rely on risk adjustment to predict health care spending, and may be using the same formulas to assess the contribution of medical conditions to overall health care spending. Typically, these formulas are estimated with parametric linear regression. The introduction of machine learning approaches has the potential to provide both improved prediction and statistical inference. I will discuss the implementation of ensembling for plan payment risk adjustment, possibly allowing for a simplified formula, thereby reducing incentives for increased coding intensity and the ability of insurers to "game" the system with aggressive diagnostic upcoding. Additionally, I will present an evaluation of how much more, on average, enrollees with each medical condition cost after controlling for demographic information and other medical conditions using double robust techniques with ensembles. Results indicate that the health spending literature may not be capturing the true incremental effect of medical conditions, potentially leaving undesirable incentives related to prevention of disease.

13.5 INVITED - Spatial Statistics: Functional data analysis for earth observation

Thursday 8 September 2pm – 3.20pm

Globolakes; Functional Clustering of MERIS lake water quality data using adaptive smoothing

Ruth O'Donnell, Claire Miller, Marian Scott
University of Glasgow

As a result of their sensitivity to climatic conditions, lakes have been described as sentinels of change. Recent developments in Earth Observation (EO) instruments such as MERIS (Medium-Spectral Resolution, Imaging Spectrometer) and AATSR (Advanced Along-Track Scanning Radiometer) from the European Space Agency's Envisat satellite platform have enabled the retrieval of a range of water quality determinands from lakes. This has resulted in the availability of expansive spatiotemporal datasets which simultaneously facilitate global assessment of environmental change and present new statistical challenges. GloboLakes is a 5 year UK (NERC) funded research programme investigating the state of lakes and their response to climatic and other environmental drivers of change. This project is investigating the coherence of water quality of lakes using EO data, at both a global and within lake scale. A functional data analysis approach has been taken with considerations for dealing with missing data and discontinuities in time series. Functional clustering methods have been also been developed to identify clusters of lakes which are similar in terms of both trends and seasonal patterns.

13.5 INVITED - Spatial Statistics: Functional data analysis for earth observation

Thursday 8 September 2pm – 3.20pm

Functional data methods for uncertainty analysis of atmospheric measurements

Francesco Finazzi, Alessandro Fassò
University of Bergamo

The Gap Analysis for Integrated Atmospheric ECV CLImate Monitoring (GAIA-CLIM) project is a Horizon2020 research project which aims to improve the ability to use ground-based and sub-orbital observations to characterise satellite observations for a number of atmospheric Essential Climate Variables (ECV).

A goal of the project is to improve understanding and uncertainty quantification of vertical profile measurements related to the same variable but collected using different instruments at different points in space and time.

In this talk, we describe a statistical modelling approach capable of explaining the relationship between collocation uncertainty and a set of environmental factors. The approach is based on the heteroskedastic functional regression model which extends the standard functional regression approach and allows a natural definition of uncertainty profiles. Along this line, a decomposition of the total collocation uncertainty is proposed, giving both a profile budget and an integrated column budget.

The approach is applied to GAIA-CLIM project data related to remote sensing and radiosonde measurements of ECVs temperature and water vapour, collocated in space and time over Europe.

13.5 INVITED - Spatial Statistics: Functional data analysis for earth observation

Thursday 8 September 2pm – 3.20pm

Trends in stratospheric ozone profiles using functional mixed models

Serge Guillas, A Park, I Petropavlovskikh
University College London

This talk is devoted to the modeling of altitude-dependent patterns of ozone variations over time. Umkehr ozone profiles from 1978 to 2011 are investigated at two locations: Boulder (USA) and Arosa (Switzerland). The study consists of two statistical stages. First we approximate ozone profiles employing an appropriate basis. To capture primary modes of ozone variations without losing essential information, a functional principal component analysis is performed. It penalizes roughness of the function and smooths excessive variations in the shape of the ozone profiles. As a result, data-driven basis functions (empirical basis functions) are obtained. The coefficients (principal component scores) corresponding to the empirical basis functions represent dominant temporal evolution in the shape of ozone profiles. We use those time series coefficients in the second statistical step to reveal the important sources of the patterns and variations in the profiles. We estimate the effects of covariates – month, year (trend), quasi-biennial oscillation, the solar cycle, the Arctic oscillation, the El Niño/Southern Oscillation cycle and the Eliassen–Palm flux – on the principal component scores of ozone profiles using additive mixed effects models. The effects are represented as smooth functions and the smooth functions are estimated by penalized regression splines. We also impose a heteroscedastic error structure that reflects the observed seasonality in the errors. The more complex error structure enables us to provide more accurate estimates of influences and trends, together with enhanced uncertainty quantification. Also, we are able to capture fine variations in the time evolution of the profiles, such as the semi-annual oscillation. We conclude by showing the trends by altitude over Boulder and Arosa, as well as for total column ozone. There are great variations in the trends across altitudes, which highlights the benefits of modeling ozone profiles.

14.1 Contributed - Methods & Theory: Multivariate Problems

Thursday 8 September 3.50pm – 4.50pm

Centering and the Use of Additional Information When Calibrating in the Presence of Random Effects

Samuel Oman
Hebrew University of Jerusalem

In calibration problems we use a calibration set, comprising values of both x (a precise measurement of a quantity of interest) and Y (an imprecise measurement which is, however, less expensive or more easily obtained), to estimate the relation between the variables. At the future, prediction step, we only observe Y and wish to estimate ξ, the corresponding unobserved x-value. Oman (1998, Biometrika 85, 439-449) examined the case where Y and x are linearly related and the calibration data consists of clustered observations obtained from different sampling units such as subjects, experiments or machines. He proposed an estimator ξ^* which accounts for subject-specific random effects when estimating ξ for a new sampling unit, as well as an estimator ξ^a for the case where we can use additional information, in the form of one or more (x, Y) observations for the new unit, to estimate subsequent x-values when only Y is obtained. Here, we extend these results in two directions. We first show that if ξ is centered about a particular value c (for example, the mean of the calibration x-values, or a value whose detection is particularly important), then ξ^a now shrinks towards c and can give substantially improved predictions in a large neighborhood of c. Second, we study the numerical performance of the estimator ξ^a, and show that as little as one additional (x, Y) observation can dramatically decrease the prediction error. We illustrate using data from a urodynamic clinic. At the calibration step, each of 23 women had up to 8 known bladder volumes x induced by catheterization, and corresponding trans-vaginal ultrasound measurements Y were obtained. At the prediction step, the objective is to use ultrasound to estimate a new subject's bladder volume, for example before and after surgery for urinary incontinence.

14.1 Contributed - Methods & Theory: Multivariate Problems

Thursday 8 September 3.50pm – 4.50pm

New multiple testing procedures for discrete test statistics

Alex Lewin, Elena Kulinskaya
Brunel University London

Commonly used multiple testing procedures controlling the Family Wise Error Rate (FWER) or the False Discovery Rate (FDR) can be conservative when used with discrete test statistics. We propose fuzzy multiple comparison procedures which give a fuzzy decision function, using the critical function of randomised p-values. We further define a new non-randomised, non-fuzzy decision rule. The new procedures are valid for a wide class of stepwise FWER and FDR controlling procedures. We also define adjusted p-values for the new multiple comparison procedures. The method is demonstrated on data sets from bioinformatics involving discrete statistics.

Kulinskaya, E. and Lewin, A. (2009). On fuzzy FWER and FDR procedures for discrete distributions. Biometrika 96, 201 – 211.

Geyer, C. and Meeden, G. (2005). Fuzzy and randomized Confidence intervals and P-values. Statistical Science 20, 358–366.

14.2 Contributed - Medical Statistics: Survival Analysis

Thursday 8 September 3.50pm – 4.50pm

Multi-Parameter Regression Models and Non-proportional Hazards

Kevin Burke, Gilbert MacKenzie
University of Limerick, Ireland

Background

The Proportional Hazards (PH) model is the most popular regression model for survival data; however, non-PH effects are often encountered in practice. One common explanation for non-PH effects is the presence of unobservable subject-specific random effects. In this so-called "frailty" model, covariates may have proportional effects at the individual level but are non-proportional at the population (i.e., marginal) level. An alternative explanation is that covariates truly exhibit non-PH effects at the individual level (and, hence, at the population level). We introduce the "Multi-Parameter Regression" (MPR) model: a flexible parametric survival model which handles this situation.

Methods

Parametric regression models typically relate covariates to one distributional parameter of specific interest, for example, in generalized linear models the location parameter is regressed on covariates while other parameters (e.g., dispersion) are constant. A more flexible approach is to regress multiple parameters simultaneously on covariates; we refer to this practice as "Multi-Parameter Regression" (MPR). We explore the Weibull MPR model (scale and shape regression components) in terms of its hazard ratio and compare with the frailty model. We also consider variable selection in the MPR model via simulation studies and real data.

Results/Conclusions

The MPR approach directly generalises the parametric PH model to non-PH (i.e., time-dependent effects) status and provides a new test of proportionality. This flexible regression model can produce dramatic improvements in fit compared to the basic PH model and its frailty extension. Combining the MPR and frailty approaches leads to a more general approach still. Interestingly, this MPR-frailty model outperforms the MPR model in the setting of our real data example showing that although both MPR and frailty approaches allow for non-PH effects, one does not abolish the need for the other.

Reference

Burke, K. and MacKenzie, G. (submitted). Multi-Parameter Regression Survival Models, Biometrics

14.2 Contributed - Medical Statistics: Survival Analysis

Thursday 8 September 3.50pm – 4.50pm

Flexible parametric survival models and time-series analysis for extrapolating time-varying treatment effects in health technology assessment.

Benjamin Kearns, Jim Chilcott, Sophie Whyte
The University of Sheffield

Objective

Health technology assessment (HTA) aims to evaluate if a health technology represents value for money. This requires estimation of the technology's effectiveness and its potential impact over a patient's lifetime. Direct evidence on time-to-event outcomes is usually only available for limited time horizons, requiring extrapolation to obtain estimates of lifetime effects. Often standard parametric models are used; for time-varying treatment effects, flexible Royston-Parmar (R-P) models may be more appropriate, but are seldom used in HTA. A case-study applies R-P models to estimate the lifetime effectiveness of screening for ovarian cancer on mortality.

Methods.

Individual patient-level data was reconstructed using published Kaplan-Meier curves. R-P models were used to estimate the time-varying hazard ratio for screening compared to no screening. Extrapolation of the hazard ratio used exponential smoothing time-series methods. Results were compared with those obtained from extrapolating standard parametric survival models (exponential, Weibull, Gompertz, log-logistic and log-Normal). Goodness of fit was evaluated using both the Akaike and Bayesian Information Criteria (AIC and BIC).

Results.

After a median of 11.1 years follow-up, observed ovarian cancer mortality was 0.34% and 0.29% amongst the no screening and screening groups. Lifetime mortality extrapolated from the R-P model was 1.84% and 0.83% respectively. The Weibull, log-logistic and log-Normal all demonstrated equally good fits, with differences in AIC and BIC of less than 0.17%. Estimated lifetime mortality ranged from 1.47% to 1.96% for no screening and from 1.00% to 1.20% for screening. Both the R-P model and the best-fitting parametric models showed similar goodness of fit with AIC of 7,453 and 7,458, and BIC of 7,532 and 7,495 respectively.

Conclusions.

The use of R-P models and time-series analysis offers a flexible approach to extrapolating treatment effects. Different plausible models can result in markedly different extrapolations; it is important that this structural uncertainty is captured within HTA.

14.2 Contributed - Medical Statistics: Survival Analysis

Thursday 8 September 3.50pm – 4.50pm

An innovative permutation approach to testing for survival curves

Luigi Salmaso, Fortunato Pesarin, Roberto Fontana, Rosa Arboretti
University of Padova

This paper deals with nonparametric inferential aspects in the field of survival analysis and particularly it is focussed on the comparison of survival curves. At first we review some recent proposals for the comparison of survival curves and related critical drawbacks. Then we propose an extension of the NonParametric Combination methodology (NPC, Pesarin F., Salmaso L. (2010) Permutation tests for complex data: theory, applications and software. John Wiley and sons) to compare survival curves by taking into account informative and non-informative censoring. Most of traditional inferential methods in survival analysis often require large sample size while, in practice, researchers have to deal with few subjects and possible many endpoints when not only mortality if of interest. On the other hand, permutation NonParametric Combination (NPC) tests represent an appealing alternative since they are distribution-free and allow for quite efficient solutions when the sample size is low. Moreover, our solution allows to test also the censoring mechanism to determine if treatment effect might also affect also the censoring along with mortality of other endpoints of interest. A peculiar advantage of our methodology consists on the possibility of immediately extending the solution to the multivariate situation taking into account of the dependency among different endpoints. A comparative simulation study is also performed to show the behaviour of the proposed methodology by comparing it to other proposals from the recent literature.

14.3 Contributed - Social Statistics: Identification and estimation

Thursday 8 September 3.50pm – 4.50pm

A general three-step method for estimating the effect of multiple latent categorical predictors on a distal outcome

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

Latent class analysis (LCA) is widely used to derive categorical variables from multivariate data which are then included as predictors of a distal outcome. The traditional 'modal class' approach is to assign subjects to the latent class with the highest posterior probability. However, regression coefficients for the modal class will be biased due to potential misclassification and the unintended influence of the distal outcome on class membership. To address these problems, a 3-step method was proposed by Asparouhov and Muthén (2014) in which the modal class is treated as an imperfect measurement of the true class in the regression for the distal outcome, with measurement error determined by the misclassification probabilities. However, their approach considers only a single latent categorical variable. This paper extends their proposition to the multiple latent categorical variable case and assesses the relative performance of the 3-step method against the traditional modal class approach under different settings. An approach based on multiple pseudo class draws (Bandeen-Roch et al. 1997) is also considered.

Simulation studies are performed using two latent categorical variables as an illustration. Models with continuous and nominal distal outcomes are estimated, under settings of associated and independent latent class variables at different entropy levels. Current results show that the 3-step method is robust and outperforms the modal class approach in most situations. In particular, when class separation is low (entropy lower than 0.8) and latent categorical variables are associated, the modal class approach produces heavily biased estimates and low confidence interval coverage. Only when entropy level is at least 0.9 do the modal class and 3-step approaches yield comparable results with low bias. The results are particularly useful for empirical studies that have more than one, possibly associated, latent constructs with unclear class separation.

14.3 Contributed - Social Statistics: Identification and estimation

Thursday 8 September 3.50pm – 4.50pm

Age-period-cohort analysis: can statistical methods solve the identification problem?

Andrew Bell, Kelvyn Jones
University of Sheffield

Age, period and cohort (APC) represent three ways in which change can occur over time: people age, time passes, and generations differ. However, these three effects, whilst conceptually distinct, are exactly mathematically collinear (since age=period-cohort). This means that attempts to model all three rely on some kind of identifying assumption, and this assumption has a huge effect on the results that are found. Despite this, methodologists have continued searching for a statistical, mechanical solution to this identification problem. Recently, Yang and Land have suggested using a cross-classified multilevel model, treating age as a polynomial fixed effect, and period and cohort groups as random effects. This succeeds in breaking the collinearity between APC; however, we have shown with simulations in previous work that the 'Hierarchical APC model' (HAPC) only produces meaningful results under specific and often unrealistic situations. Yang, Land and others have questioned this, and the debate is continuing.

This talk will summarise the debate so far, and then extend it to our current research, that moves away from using simulations, and instead uses real data. It is argued that the results found in many prior studies are the result of the inherent structures of the data, and not the substantive APC processes occurring in society. We will reanalyse a dataset used by Yang, Land and others in the past, and see if there is an effect of changing the data structure of this dataset on how the HAPC model assigns APC trends. This is instructive not just in showing whether the HAPC model is problematic, but also in revealing why it produces the results that it does.

The talk will conclude with recommendations for applied researchers looking to find age, period, and/or cohort effects in a robust manner, specifically by making informed and clearly stated assumptions about APC.

14.3 Contributed - Social Statistics: Identification and estimation

Thursday 8 September 3.50pm – 4.50pm

Producing and interpreting estimates of the prevalence of female genital mutilation in England and Wales

Alison Macfarlane, Efua Dorkenoo, Alex Kachkaev
City University London

Background

Female genital mutilation (FGM) is practised in 29 countries in Africa and the Middle East where national data are collected about it. FGM has also been documented in specific populations in other parts of the Middle East and Asia, including Indonesia, Malaysia, India, Pakistan, Oman, Saudi Arabia and the United Arab Emirates. Increasingly women born in these countries have migrated to England and Wales. Estimates of its prevalence are needed in order to plan health care for affected women and, where necessary, preventive care for their daughters.

Methods

Proxy estimates of age specific prevalence from standardised surveys undertaken in the 29 countries were applied to the numbers of women born in those countries who were enumerated in the 2011 census or who registered births in England and Wales to estimate the numbers of affected women, the numbers of affected women giving birth and the numbers of daughters born to affected women living in each local authority area in England and Wales. Individual anonymised census and birth registration records were analysed which made it possible to exclude women from religious and ethnic groups who were unlikely to practise FGM. The estimates were published in tables and interactive maps.

Results

The highest estimated prevalence rates were in London boroughs, with the highest being 4.7% of women in Southwark and 3.9% in Brent, compared to 0.5% in England and Wales as a whole. Outside London, Manchester, Slough, Bristol, Leicester and Birmingham have high prevalence rates, ranging from 1.2 to 1.6%. Nowhere had a prevalence of zero, so affected women and girls are likely to be living in every local authority area.

Interpretation

Care is needed in interpreting these estimates for many reasons. It is not possible to quantify the prevalence or risk of FGM among girls born in England and Wales.

14.6 INVITED - Significance Writing Competition Winners

Thursday 8 September 3.50pm – 4.50pm

Queen Elizabeth II - an Extreme Event monarch?

Anastasia Frantsuzova
City University/St Andrews

This year celebrates Queen Elizabeth II stepping into the tenth decade of her life and continuing her role as the longest-reigning British monarch in history. Many of us will have been enjoying the festivities of the occasion, and as statisticians we can add much more than a pie chart to the street party snacks selection. I will introduce a statistical field illustrated perfectly by Her Majesty - Extreme Value Theory. Using this approach, I will model the length of reign of monarchs from 2700 B.C to today, with a particular interest in longest-reigning rulers. First, I will assume stationarity of data, and then include possible appropriate covariates and discuss their relevance.

14.6 INVITED - Significance Writing Competition Winners

Thursday 8 September 3.50pm – 4.50pm

On the frequency of America in America

Adam Kashlak
University of Cambridge

With the upcoming US presidential election this November, political rhetoric is on the rise. While there is much uncertainly concerning its outcome, when the newly elected POTUS 45 addresses a joint session of congress in early 2017, one thing is known for certain: the new American president will speak frequently about America. To determine how frequently, we will examine historical data from present day Barack Obama to the founding father George Washington who in 1790 first addressed congress in what has now come to be known as the State of the Union (SOTU) address. Across the 227 years' worth of SOTU speeches and written reports, we will track the usage of one specific word: "America."

14.6 INVITED - Significance Writing Competition Winners

Thursday 8 September 3.50pm – 4.50pm

How to mend a broken heart with stem cells and discrepancies?

Hakim-Moulay Dehbi
Imperial College London

Imagine that you discover, by chance, *fractional* patients, and other disconcerting phenomena, in a series of reports on stem cells trials. You read for instance, that 50% of 9 patients are taking a drug. You also read that in a trial of 41 patients with two treatment arms, each arm has 21 patients. How would you react?

A group of scientists at Imperial College London, which I was part of, was in this very situation back in 2013. We discovered more than 200 discrepancies in 49 reports from one cardiology laboratory. Our curiosity was aroused. We decided to scrutinize the whole field of stem cells trials in cardiology for discrepancies. What we found certainly created much heartache, and it became much less certain that stem cells can actually mend a broken heart.

14.7 Contributed - Communicating Statistics: E-learning

Thursday 8 September 3.50pm – 4.50pm

Teaching statistics to non-statisticians on a fully online master of public health programme: past, present and future

Isla Gemmell
University of Manchester

This presentation describes the development and teaching of a Biostatistics module within a fully online distance learning Master of Public Health (MPH) programme at the University of Manchester. All materials are delivered online via the Blackboard virtual learning environment. Students on the programme are based in over 40 countries worldwide and come from a wide variety of professional backgrounds. Most students study part-time while remaining in their own country and in their existing employment.

We have developed teaching materials in line with the online learning pedagogic approach and utilised current technologies to allow us to enhance the learning experience for all our students. This has been carried out within a dedicated multi-disciplinary team incorporating academics and e-learning technologists. We will explore current practices in teaching and assessment within the module and discuss future developments in teaching statistics online within the context of the growth of online distance learning globally and the emergence of new technologies.

Finally we will describe the results of a number of our pedagogic research projects that explore; staff attitudes to online distance learning and students experience, attainment and use of online learning support materials within the programme.

14.7 Contributed - Communicating Statistics: E-learning

Thursday 8 September 3.50pm – 4.50pm

Creating a statistical analysis assistant using Stat-JR

William Browne, Richard Parker, Chris Charlton, Danius Michaelides, Luc Moreau
University of Bristol

The process of statistical analysis can be complicated with each research hypothesis and dataset posing it's own unique challenges. Historically the more complicated the analysis the more the need for the applied researcher to consult and/or collaborate with a statistician. With the explosion in both availability and quantity of datasets and the people pipeline problem of not enough statisticians to satisfy the growing population of applied researchers requiring such consultation other solutions should be researched. At the Centre for Multilevel Modelling (http://www.bristol.ac.uk/cmm/) we have specialised over the years in producing user-friendly statistical software e.g. MLwiN for particular more advanced statistical models such as multilevel models. Our software has been coupled with large amounts of documentation and training materials e.g. our LEMMA training course (http://www.bristol.ac.uk/cmm/learning/online-course/) which has had nearly 20,000 users.

In our recent ESRC funded research we have been looking at how we might translate our research from producing training materials into the creation of a statistical analysis assistant. This is a computer program that for particular types of statistical analysis will prompt the user for inputs and then perform the standard steps of such an analysis while explaining to the user what it is doing and why thus teaching the user the techniques while using their own dataset. The output of the program will be an annotated analysis that can then be shared and this will aid in reproducible research. In this talk we will describe how such an analysis assistant works for analyses including linear models, multilevel models, logistic regression, MCMC estimation and dealing with missing data.

14.7 Contributed - Communicating Statistics: E-learning

Thursday 8 September 3.50pm – 4.50pm

Taking the sting out of stats: Teaching and communicating stats effectively for e-learners

Eirini Tatsi, Sophie Drennan
University of Derby, Online Learning

In the last decade, the use of computer-mediated communication (CMC) has been widely adopted as a primary delivery method for online education (Mills & Raju, 2011). Teaching an online statistics course thus poses more of a challenge, both for the academic and the learners alike due to the levels of student preparedness, the perceived complexity of the information and the potential lack of 'hands on' tutorial opportunities.

At UDOL, we strongly believe that interactivity and visualisation of the materials are aspects which differentiate an effective course/module from a correspondence course. In order to make sure that our students receive high quality teaching via an intellectually stimulating environment UDOL has responded by developing a variety of visual and interactive techniques within the online learning environment with which to promote a more accessible learning format for statistical analyses. Such techniques specifically aim to alleviate the anticipatory 'fear' often accompanying statistical analyses and offer students the opportunities of authentic practical application and 'real-time' tutorials.

As a result, the qualitative and quantitative feedback gained from students following the implementation and inclusion of these specifically tailored techniques has been encouraging: they gain understanding of statistical strategies and therefore feel confident in achieving their goals and completing their courses successfully.

This presentation aims to provide information on how we teach and communicate statistics in a virtual learning environment. It will offer examples of the primary techniques involved and discuss the implications of the benefits to our online students in creating an understanding of statistical analyses.

Plenary 5 – Significance Lecture

Thursday 8 September 5pm – 5.50pm

Lost in translation – Why statisticians and policymakers need to speak the same language

Dame Anne Glover
University of Aberdeen

Statisticians frequently meet with policy makers to discuss data and evidence. They talk, they listen – but does either side truly understand the other? Drawing on her past experience as Chief Scientific Adviser to the President of the European Commission and Chief Scientific Adviser for Scotland, Professor Dame Anne Glover will discuss the role of science and evidence in policy making, the value of collaboration and clear communication, and the vital importance of statistical education.

POSTER PRESENTATIONS

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

Efficiency of Some Exponential Estimators for Estimating Heterogeneous Population Parameters

Olaniyi Mathew Olayiwola, Timothy Olubiyi Ayeleso
Federal University of Agriculture, Abneokuta, Nigeria

Sratified Ranked Set sampling (SRSS) helps in obtaining an unbiased estimator for populat ion parameters with some significant gain in efficiency. This paper presents modified exponential estimators of finite populat ion mean using co-efficient of Variat ion and Co-efficient of Kurtosis of auxiliary variable. The bias and mean square error (MSE) of the proposed estimators with large sample approximat ion were derived. A set of secondary data on students' enrolment in secondary schools in Ogun State were used. The population was stratified into 4 strata base on political zones which are Egba, Yewa, Ijebu and Remo with 89, 91, 69 and 53 schools respectively. The sample sizes of the 4 strata based on proportional allocation are 27, 27, 21 and 15 schools respectively. The population means for students' enrolment and number of staff are 1284.71 and 49.0 respectively. The students' enrolment in Egba, Yewa, Ijebu and Remo zones are 140718, 137835, 56618 and 52815 respectively. The proportions of staff members to number of enrolled students are 0.042, 0,025, 0.051 and 0.047 respectively. The MSEs for four proposed estimators are 6176.84, 6503.61, 6269.63, and 6632.94. The MSEs for the corresponding four existing estimators are 9754.51, 10270.80, 9748.76, and 10561.44. The proposed estimators have least MSE, hence they are more efficient

Spatial Patterns of Tuberculosis Incidence in Buffalo City Municipality, Eastern Cape Province, South Africa: The Disease Map Approach

Davies Obaromi, Ngozi Johnson, Yongsong Qin, James Ndege
Department of Statistics, University of Fort Hare, South Africa.

OBJECTIVE: The purpose of this research was to investigate the spatial and temporal distribution of tuberculosis (TB) during 2008 to 2015, and to also identify clusters of higher prevalence and incidence of TB in Buffalo City Municipality, Eastern Cape Province in South Africa, in order to generate a disease atlas.

METHODS: Spatial patterns and distribution of disease incidence were investigated using the ArcGIS 9.2 and descriptive statistics by a graphical ranking method imposed on the map. Also clustering techniques was used to identify the exact locations for high and low incidences of TB. The cluster mappings were also produced as building blocks in the profiling of the TB cases. GIS queries were also carried out to further emphasize the clustering patterns of the disease for the period of 2008-2015.

Results: From our investigations and findings, four clusters of high incidence of TB was identified in East London, Mdantsane, Duncan village and King William's Town for all the years under study. This high prevalence of TB cases in those locations also showed a high association with HIV/AIDS prevalence. The GIS showed an eastward outlook in the prevalence of TB in the municipality. The combined cluster map and GIS queries also showed an eastward high prevalence of TB.

CONCLUSION: TB prevalence in the municipality showed a systematic pattern in the distribution of the disease cases in the region and is found to concentrate in areas of high HIV/AIDS rates and also areas mostly populated by Blacks. Multifaceted and hidden relationships may exist between TB incidence and a wide range of environmental and inherent factors, which call for future research.

Analysis of Nigeria Stock Exchange Allshare Index

Stephen Adebesin
Yaba College of Technology

The study focus on time series analysis of All-share index of Nigerian Stock Exchange from the period of (January 1985 – December 2014). The data employed for this study is secondary data gotten from Nigerian Stock Exchange (NSE).The data was analyzed using R Software version 3.1.3 and the statistical tool employed is Time Series Analysis using Box-Jenkins approach. A sequence plot of the observations showed underlying variation which makes the data non- stationary and transformation was made by differencing once before the stationarity of the data was obtained. An ARIMA model was employed to forecast All-share index. Hence the model fitted for the above is (1, 1, 2) and can be represented mathematically as $\mathbf{ASI_t = 0.0702ASI_{t-1} - 1.0462\varepsilon_{t-1} + 0.0463\varepsilon_{t-2}}$. In addition, the Box-Pierce chi- square p-value 0.9287 is not-significant indicating that the model fits the data well. The forecasted value fitted by ARIMA (1,1,2) indicate that All-share index will decrease for the first six period of year 2015 and the remain constant for the next periods up of year 2016 and 2017.We therefore recommend that there is the need for government to implement prudent macroeconomic policies in order for the country to derive maximum benefits from the Capital Market.

A review of outcome measures in in vitro fertilisation (IVF) randomised controlled trials.

Jack Wilkinson, Stephen Roberts, Marian Showell, Andy Vail
University of Manchester

Study question: How are outcome measures reported in randomised controlled trials (RCTs) for in vitro fertilisation (IVF)?

Summary answer: Many combinations of numerator and denominator are in use, and are often employed in a manner that compromises the validity of the study.

What is known already: The choice of numerator and denominator governs the relevance and statistical integrity of a study's results.

Study design, size, duration: Review of outcomes reported in 142 IVF RCTs published in 2013 or 2014.

Participants/materials, setting, methods: Trials were identified by searching the Cochrane Gynaecology and Fertility Specialised Register. Reported numerators and denominators were extracted. We checked to see if live birth rates were calculated using the randomised cohort or another denominator.

Main results and the role of chance: Over 800 combinations of numerator and denominator were identified. No outcome appeared in a majority of trials. Twenty-two (43%) of studies reporting live birth presented a calculation including all randomised participants or excluding small numbers of protocol violators.

Limitations, reasons for caution: Several of the included articles may have been secondary publications. Our categorisation scheme was essentially arbitrary, so the frequencies we present should be interpreted with this in mind. The analysis of live birth denominators was post-hoc.

Conclusions: This review suggests large-scale inconsistency in outcome reporting and widespread misunderstanding of RCT methodology in IVF research. This presents a barrier to the synthesis of results in meta-analysis and to the evaluation of infertility interventions.

Measuring the Outcomes of the Portuguese Nursing Homes and Community-Based Services' Populations

Hugo Lopes, Nicoletta Rosati, Céu Mateus
National School of Public Health (Escola Nacional de Saúde Pública, Portugal)

Objectives: The main goals for this work are to identify the differences between patients' characteristics in three Portuguese Nursing Homes (NH) typologies and Home and Community-Based Services (HCBS) settings of care, evaluate the differences between their cognitive and physical outcomes, and identify the patients' characteristics which mostly influence the outcomes in each setting of care.

Methods: Chi-square and Mann-Whitney tests were used to identify differences between all populations at study; Wilcoxon signed-rank test was performed to quantify the number of patients with a dependency level at discharge lower, higher or equal to their dependency level at admission in two areas (cognitive and activities of daily living). In order to analyse the difference between the mean scores from admission and discharge in the several activities assessed in each area, was performed a paired-sample t-test. Finally, to identify the patients' characteristics at admission which better influence their outcomes in each setting of care, Binaries regressions were performed using the Forward:LR method.

Results: The HCBS population was less cognitively impaired but more physically dependent than NH population. Regarding the differences between dependency levels from admission and discharge, although a higher percentage of HCBS population improved their cognitive status, a higher percentage of patients improved their physical functions at discharge at NH setting of care. Finally, the variables which influence the probability to improve the physical independence levels in both settings are age, length of care, cognitive and physical status at admission, and so some main groups of pathologies. Nevertheless, further study needs to demonstrate the effect of the LTC rehabilitation processes to better explain and predict the patients' outcome.

Mixture of Complete Pooling, Partial Pooling, and No-Pooling Models in Bayesian Subgroup Analysis

Jun Takeda
Astellas Pharma Inc.

In clinical trials the sites or the countries can be seen as subgroups. Some oncology trials have subgroups with cancer subtypes. Modeling of these subgroups in statistical analysis is an important issue to consider. If some similarity among subgroups can be assumed, the hierarchical model, also called the partial pooling model, is typically employed in Bayesian analysis to borrow strength. However, in some cases the characteristics of data cannot be well explained with a simple partial pooling model. To deal with this issue, Neuenschwander et al (2016) proposed a mixture of the partial pooling model and the no pooling model, in which all subgroups are treated independently. In this research their mixture model was further extended to deal with data that has some nature of complete pooling, where all subgroups are assumed to have only common parameters. To explore models we also suggested the following procedure with the extended mixture model:

-Define a mixture model of complete pooling, partial pooling, and no-pooling models

-Construct a model family with the submodels

-Compare models in the family with WAIC (widely applicable information criterion)

The above procedure was applied to various data sets with subgroups, including data sets from global clinical trials, oncology trials, and meta-analysis settings. In some cases models including a component of complete pooling showed lower WAICs than other models. This means that inclusion of complete pooling can be one way to improve modeling in Bayesian subgroup analysis. In addition we discovered that comprehensive comparison among models in the family gives some insight to understand the nature of the data.

Reference

Neuenschwander B, Wandel S, Roychoudhury S, and Bailey S (2016) Robust exchangeability designs for early phase clinical trials with multiple strata. Pharm Stat. 2016 Mar;15(2):123-34

Sleep - some results from the China Kadoorie Biobank

Christiana Kartsonaki, Yiping Chen, Bian Zheng, Guo Yu, Robert Clarke, Zhengming Chen
University of Oxford

Sleep is very important for health. Some analyses on sleep in the China Kadoorie Biobank (CKB) will be presented. The CKB is a prospective cohort study of over 0.5 million Chinese adults aged 30-79 years recruited from 10 diverse regions (5 rural and 5 urban) of China between 2004 and 2008. Study participants were given a questionnaire which includes detailed demographic, socioeconomic, health-related and lifestyle factors. Physical measurements were made by trained staff at local assessment centres. A 10ml blood sample was collected and prepared for long term storage. About 5% of participants have been re-surveyed twice after entry into the study. A case-control study for cardiovascular outcomes nested within the CKB cohort has been done for various blood-based measurements. Some findings on the associations of sleep duration and other sleep characteristics with other features measured at entry into the study will be presented and methodological issues that arise will be discussed.

Who are the Portuguese Long-Term Care Beneficiaries? Differences Between Nursing Homes and Community-Based Services Populations

Hugo Lopes, Nicoletta Rosati, Céu Mateus
National School of Public Health (Escola Nacional de Saúde Pública, Portugal)

Objectives: In order to address the challenges of ageing population, the Portuguese National Network for Long-term Integrated Care (NNLIC) was created in 2006, organized in home and community-based services (HCBS) settings of care and four typologies of nursing homes (NH). The main goal is to explore, for the first time in Portugal, to what extent three NH typologies and HCBS populations differ from each other, trying to shed light in the criteria used for referral to each typology.

Methods: Based on the system used to assess the dependency levels of LTC beneficiaries, patients characteristics are compared between all typologies of care; patients' dependency levels are studied at admission in various activities evaluated in three areas (cognitive, activities of daily living and locomotion). In order to determine the probability of a patient to receive care in the two main settings of care, a logistic regression was performed. Finally, due to the fact that the three NH typologies follow an intrinsic order as the dependency level of the patients increase, the ordered logistic regression model was used to determine the main characteristics of the patients in each NH typologies.

Results: There were differences regarding sociodemographic characteristics, main groups of pathologies and dependency levels at admission between the populations of all typologies of care. Compared with those receiving care at HCBS, being female, not being married, being literate, having specific pathologies, being classified in a higher level of ADL impairment, being referred by the hospital, and a longer care responsiveness process increases the probability to receive care in a NH. According to the ordered logistic regression model, the patients correctly referred ranged between 31% and 76%. Patients classified in a higher independence cognitive and ADL status at admission, increase the probability to be referred to the short-term care typology.

A latent class growth curve model for walking behaviour in an indoor mobility test

Carla Rampichini, Leonardo Grilli
University of Florence, Department of Statistics, Computer Science, Applications 'G. Parenti'

This work is motivated by a study on aging based on a representative population living in the Chianti geographic area (Tuscany, Italy). Multiple factors may influence the ability to walk and no standard criteria are currently available to establish whether these factors are functioning within the “normal” range. Our work exploits data collected during the performance of an indoor mobility test to discriminate individuals at high-risk of mobility disability or falls. A large number of outcomes was collected during the test. We specify a latent class growth curve model to detect walking impairment. The model explicitly considers non-ignorable missing data. The effect of subject pre-test characteristics on the probability of belonging to each cluster is also investigated. Our findings suggest that subjects aggregate into distinct behavioural profiles, characterized by age and other demographic and anthropometric factors.

Estimating equation projecting estimators for additive hazard model with missing covariates

Shuqin Fan, Alan Wan, Yong Zhou
University of Chinese Academic Sciences & City University of Hong Kong

Biomedical studies in missing covariate are common these days, especially in the additive hazards regression model with some missing at random covariates. However, those estimating methods need to estimate the selection probability before give the final estimators of the regression parameters. In this paper, we propose a method, called estimating equation projecting(EEP) method, to estimate the regression parameters without estimating the selection probability. In order to complete our method, we estimate the missing covariates in both parametric and nonparametric way. The resulting estimators are shown to be consistent and asymptotically normal. Also we add the nonparametric estimated missing probability and missing covariates in augmented weighted methods, which nobody has done before. Simulation studies show that the proposed EEP estimators with parametric and nonparametric estimators of missing covariates slightly outperform the augmented weighted estimators in both biases and variances. An application to the mouse leukemia data is provided to see how it performs in reality.

Modelling conditions and health care processes in electronic health records through clinical codes

Ivan Olier, Evangelos Kontopantelis, Chun Shing Kwok, Darren Ashcroft, Mamas A Mamas
Keele University

Background

The use of Electronic Health Records databases is becoming more commonplace for medical research. Electronic Health Records research often begins with the development of a list of clinical codes with which to identify cases with a specific condition. We present a methodology and accompanying Stata/R commands, (*pcdsearch)*, to help researchers in this task. We use Coronary Heart Disease (CHD) and Heart Failure (HF) as example conditions.

Methods

We used the Clinical Practice Research Datalink, a UK Primary Care Database in which clinical information is largely organised using Read codes, a hierarchical clinical coding system. *pcdsearch* is used to identify potentially relevant clinical codes and/or product codes from word-stubs and code-stubs suggested by clinicians. The returned code-lists are reviewed and codes relevant to the condition of interest are selected. The final code-list is then used to identify patients.

Results

We identified 169 and 99 Read codes linked to CHD and HF, respectively, and used them to identify cases in the database. We observed that our approach identified cases that would have been missed with a simpler approach using CHD and HF registers defined within the UK Quality and Outcomes Framework.

Discussion

We described a framework for researchers using Electronic Health Records databases to identify patients with a particular condition or matching certain clinical criteria. The method is applicable across different coding systems and databases, and can be used with SNOMED CT, ICD coding or other medical classification code-lists.

Conclusion

The code selection process is an important methodological component in analyses of routinely collected data and needs careful consideration.

Building Econometric model by using open data statistics for the determination of Oil Prices”

Thamer Zaidan
League of Arab States/ Egypt

Results which look good at first sight may be riddled with problems if we care to look at our data statistics of oil prices more carefully. Many of the problems which emerge at the level of multivariate analysis can often be traced back to particularities of the data statistics.

OAPEC open data statistics is free and open access to data of oil. Transparency is central to challenging the " Recourse Curse" which has left many countries with substantial natural resource wealth among the poorest Over the coming years, a flood of new data on oil prices is anticipated, as new rules come into force around the world. This creates exciting new challenges ahead to equip the oil sector to make the most of this data, while also filtering out many of the endogenous factors that have historically contributed to change in oil prices.

The most important factors affecting the price of oil are , developments in supply and demand, World Economic Growth, and other economic factors coincided with geopolitical factors and speculations, climate factors, scarcity factors, technical factors, cash factors , the production of unconventional oil .

This paper aims:
1. To draw upon modern approaches to data analysis of oil prices.
2. To make extensive use of graphical methods in data statistics analysis.
3. to build an econometric model to indicate the independent variables explain the behavior of oil prices , during 1995-2015.

To achieve these objectives, the following questions will be asked:
1. Which is the most appropriate variants of modern approaches to data statistics analysis of oil prices ?
2. Which are the most important variables affect the price of oil in order to give some insight on where to focus to control the price of oil?

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

Khaled Alqahtani, Arief Gusnanto, Charles Taylor
University of Leeds

Non-small-cell lung cancer is one of the main sources of death around the world. As a result, scientists are now looking for some of the risk factors for lung cancer which can be caused by certain changes in the DNA of lung cells (CNA). Next-generation sequencing (NGS) technologies produce high-dimensional data that allow a nearly complete evaluation of genetic variation which makes the number of covariates are greatly exceeds the number of observations.

The results of our analysis indicate that we can incorporate the copy number alteration profile to predict the survival time. We introduce a novel algorithm based on smooth extended cox model (SCOX) within a random effects model-frame work using penalized partial likelihood (PPL) to model the survival time based on the patients' clinical characteristics as fixed effects and CNA profiles as random effects. We assumed CNA coefficients to be correlated random effect that follow a mixture of tow distribution to imposed smoothness. In order to deal with the sudden jump or declined, we assume one of the mixture distribution is Cauchy distribution. The other mixture is assumed to follow a normal distribution. SCOX method does not automatically lead to selection of relevant variables because SCOX construct a linear combination of all original predictors. Therefore, We also propose a new method of sparse SCOX (SSCOX) for survival data to allow sparse variable selection, smoothness, and dimension reduction in the same time. This can bee achieved by assuming CNA coefficients to be correlated random effect that follow a mixture of three distribution:Normal, second difference of Cauchy to achieve smoothness, Laplace to achieve sparsity. For the tuning parameter estimates we used cross-validation partial likelihood . We find that the our models is suitable and has enabled a survival probability prediction based on the patients' clinical information and CNA profiles.

Bayesian Proportional Hazards and insights from censored quantile regression model for paediatric and adolescent HIV/AIDS patients on antiretroviral treatment.

Innocent Maposa, Renette Blignaut
Namibia University of Science and Technology

Survival analysis techniques are often used in biostatistics, epidemiological and clinical research to model time until event data. These techniques have recently been enhanced through modern computational advancements.

The purpose of this study is to fit a bayesian proportional hazards model and a censored quantile regression model to paediatric data and then compare the results in terms of inferences on the effect of the different prognostic risk factors on ART patient survival times. A retrospective cohort study design was conducted for children who initiated anti-retroviral treatment (ART) between 01 January 2006 and 31 December 2010. A sample of 813 children was used. Imputation was performed for all variables that had missing cases.

The results from a bayesian proportional hazards model indicate that not being an infant had a positive effect on survival time. Patients initiating treatment in clinical stage II instead of stage IV had a significant positive effect on survival time. The results from the censored quantile regression model are more revealing, highlighting that initiating in clinical stageII had a significant positive effect on survival time during early periods of initiation only compared to initiating in clinical stage IV. The results also reveals that patient gender has a significant effect during the early periods of starting ART only but not significant at any other point during treatment. These insights are possible from censored quantile regression perspective.

The conclusion from this study is that more insights are obtained from using censored quantile regression models as compared to the proportional hazards models framework. However, noting that the censored quantile regression models do not give hazard ratios, it is our belief that if these models are applied together, then we can get both the important insights into the dynamics of the prognostic risk factor effects as well as the hazard ratios associated with them.

Combining sources of evidence to identify gene-disease associations as targets for drug development

Nicholas Galwey, Paul Fisher, Paul Wilson
GlaxoSmithKline

Evidence is available from a range of sources – genetics, pharmacology, scientific literature etc. – on reported associations between human genes and diseases, scored as 1 (evidence of association) or 0 (no reported evidence): genes associated with disease are potential 'targets' for pharmaceutical drug development. A weighted sum of scores from several sources of evidence can be used to rank novel gene-disease combinations in order of priority for investment of research effort, but how should appropriate weights be chosen? They should clearly reflect expert judgement of the relevance and reliability of each evidence source, but should they also reflect the information in the covariance matrix among the scores? These questions were explored using the association scores for >32,000 representative gene-disease combinations. The score for each source could be standardised according to its standard deviation (SD), and/or the weights given to different sources could be adjusted to take account of the covariances between them, in the same way that univariate regression coefficients are related to multiple regression coefficients obtained from the same data. The impact of standardisation and adjustment was assessed in an independent set of >21,000 gene-disease combinations. In the representative data set there was wide variation in the SD (range 0.048 to 0.498), so a decision to standardise had a considerable effect. However, covariances between sources were weak (range of correlation coefficients: r = -0.296 to 0.201), so a decision to adjust for covariance made little difference. The weighted sums were compared with known associations (i.e. gene-disease combinations for which a drug has been launched) in the independent data set. There was no weighting method that out-performed the un-weighted sum, as assessed by positive predictive value or relative risk. However, we recommend that users should continue attempts to specify appropriate weights based on new expert knowledge as it arises.

Efficacy Study of Two Combined Anti-hypertensive Drug in Reducing Blood Pressure Characteristics: A Multivariate Statistical Approach

Ekele Alih, Benson Ade Afere, Fidelis Ifeanyi Ugwuowo, Yunusa Maji
Federal Polytechnic Idah, Kogi State, Nigeria

Most hypertensive patients require at least two different drug classes to achieve the recommended target blood pressure of 130/85 mmHg as a single dose may not easily bring it down as fast as possible (Mahmud and Feely, 2015). This study evaluates the efficacy of two combined anti-hypertensive therapy namely: Captopril+hydrochlorothiazide, (C_h) and Lisinopril + hydrochlorothiazide, (L_h). A total of 400 patients were partitioned randomly into two groups that is: group I, (G1) and group II, (G2). G1 was given C_h while G2 was given L_h for a period of January-October, 2015 (i.e ten months). At each of the (ten) monthly visit, the systolic pressure, (SP), diastolic pressure, (DP), and the body mass index, (BMI) were recorded. Results obtained showed a reduction in blood pressure characteristics after the first visit on both groups (p=0.0004); a confirmation that hypertensive patients generally respond to two combined therapy than a single drug. There was no significant difference between the two groups at the first four visits (p_t,t=0,...,3=0.7607, 0.8279, 0.5046, 0.2208). However, a significant difference occurred from the fifth visit until the tenth visit and in favour of G2 (p= 0.0000). Barring friendliness, G2 attained the baseline at the sixth visit (p=0.0005) while G1 attained the baseline at the ninth visit (p=0.0105) and hence, it is concluded that (L_h) has a higher efficacy when compared to (C_h).

Sensitivity analysis for informative censoring in parametric survival models: An evaluation of the method

Panayiotis Bobotas, Alan Kimber, Stefanie Biedermann
University of Southampton

Siannis et al. (2005) and Siannis (2004) proposed and studied a sensitivity analysis for informative censoring in parametric survival analysis. They introduced a parametric model that allows for dependence between the failure and censoring processes in terms of a parameter δ and a bias function B(t,θ), where θ is a parameter associated with the failure process. More specifically, δ can be thought of as measuring the size of the dependence between the two processes and B(t,θ) as measuring the pattern of this dependence.

In this talk, first some theoretical issues concerning the modelling of the dependence used by Siannis et al. (2005) and Siannis (2004) are discussed. Then the results of an extensive simulation study are reported. These indicate some shortcomings of the proposed sensitivity analysis, particularly in the presence of nuisance parameters. Finally, some methods for handling informative censoring are sketched briefly.

References

Siannis, F. (2004). Applications of a parametric model for informative censoring. Biometrics*, 60, 704-714.*

Siannis, F., Copas, J., Lu, G. (2005). Sensitivity analysis for informative censoring in parametric survival models. Biostatistics*, 6, 77-91.*

Spectrum and Bispectrum of Integer Valued Time Series Models

Mahmoud El-Hashash, Mahmoud Gabr
Bridgewater State University

Integer valued time series have been the object of growing interest in recent years. One of these models is the INteger-AutoRegressive (INAR) model. Al-Osh and Alzaid (1987) introduced and studied count valued INAR models with Poisson marginals. These models are constructed based on the binomial thinning operator. This has led to the construction of integer valued ARMA (INARMA) models. In the subsequent developments of the INAR models in the statistical literature one finds a wide variety of types of such models with Poisson, Geometric, New Geometric, Binomial, and Negative Binomial marginals. R. Keith Freeland (2010) and Wagner Barreto-Souza and Marcelo Bourguignon (2013) constructed and studied true INAR(1) models, in which the observed values of the time series could be negative as well as positive.

Maria Eduarda Da Silva and Vera Lu´ cia Oliveira (2004) have calculated higher-order moments and cumulants and also the spectral and bispectral densities of the INAR(1) process with Poisson marginals while Hassan S. Bakouch (2010) has calculated them with Geometric marginals.

Different integer valued time series models may have different shapes of spectrum and bispectrum. Our goal in this study is to differentiate between the INAR processes with different distributional marginals or thinings using the spectral and bispectral densities of these models. This requires calculating the second and the third order moments and cumulants of these models with different distributional marginals and thinings using the same technique(s) that Da Silva(20014) and Bakouch(2010) have used. We also may discover some properties that characterize each model.

Design choices, integration and analysis of on-demand and on-premise data using SAP HANA cloud solutions

Dobrinka Stefanova, Rumiana Antonova
Sofia University, Faculty of Mathematics and Informatics

Nowadays, cloud computing allows companies to outsource their enterprise systems to the cloud, benefiting from the cost-effectiveness and flexibility of the cloud. Cloud vendors try to convince us of the advantages of migrating to the cloud. With large existing investments in legacy on-premise systems, however, it is very difficult for companies to make full transition to the cloud. Instead, hybrid scenarios combining on-demand and on-premise data become increasingly popular.

The current presentation shows a research of the needs for integrating enterprise data from different sources, and possible scenarios for efficient data management and analysis. It stresses on the big data processing capabilities of SAP HANA database management system, and the integration capabilities available in SAP HANA Cloud Platform.

Zero Inflated models vs Vanilla models in modelling the intensity of capture of cockroaches in a congested middle class Nigerian community.

Mary Akinyemi, Adedotun Adenusi, Bamidele Akinsanya
University of Lagos

Count data occur naturally in a number of disciplines ranging from economics and the social sciences to finance as well as medical sciences. Most count data are plagued with over-dispersion and excess zeros making it difficult to model them with vanilla linear models. Different models have been proposed to capture this peculiarity in count data viz.: Classical models such as the generalized Poisson regression model and the negative binomial regression model have been used to model dispersed count data. zero-inflated models are also said to be able to capture over-dispersion and excess zeros in count data.

Cockroaches are among the most common pests in private, public dwellings and health facilities. Their presence can raise safety concerns, especially as they maybe carriers of various pathogenic organisms. Two predominamt cockroach species were identified in this study viz: P. americana and B. germanica with the P. americana specie being indentified as the most dominant specie in most Nigerian homes.

In this paper, we compare the performance of zero-inflated Poisson and Negative Binomial models to classical Poisson and Negative Binomial regression models to modelling the intensity of coackroaches at various sites within 3 different house types in a congested Lagos community in Nigeria.

The model parameters are estimated using the method of maximum likelihood. The models' performances are compared based on their information criteria (AIC and BIC), 10-fold cross validation mean squared error, The Vuong test and Gini index. The vanilla Poisson model out performed the other models considered.

Evasion in the Brazilian Army: A Logistic Regression Model

Cleber Iack, Helena Mouriño
Faculdade de Ciências - Universidade de Lisboa / Ministério da Defesa - Brasil / Bolsista do CNPq – Brasil

Objectives

The Brazilian Armed Forces are facing a large evasion of their officers. Between 2011 and 2013, 652 officers requested early departure from the Forces, which corresponds to an increasing of 63% when compared to the period 2009-2010.

This study aims at identifying the variables intrinsically related with the early exit of the Army officers.

Methods

The sample that we studied straddle 16540 militaries between 2009 and 2014. Logistic regression was used to estimate the associations of Military school (IME, AMAN, EsSEx and EsAEx), military post and the periodic evaluations (19 items, each in a Likert-type scale) with early exit.

Results

There is a reduction of 53% in the odds of officers' leaving at the Post of Captain (OR 0.47, 95% CI 0.38-0.59), when compared to Lieutenant. Concerning the Major's post (OR 0.05, 95% CI 0.03-0.08), there is a reduction of 95% (compared to Lieutenant).

Officers from the EsSEx (OR 0.38, 95% CI 0.28-0.51), presents 62% less chances of leaving prematurely when compared to the IME. For the AMAN (OR 0.11, 95% CI 0.09-0.15), and EsAEx (OR 0.09, 95% CI 0.06-0.14),, there is a reduction of 89% and 91%, respectively (compared to the IME).

Militaries that have higher evaluation in Military Attitude (OR 0.41, 95% CI 0.26-0,63), Military Posture (OR 0.49, 95% CI 0.32-0,76) or Physical Endurance (OR 0.66, 95% CI 0.44-0.98), have less chance of leaving early. Militaries who have higher evaluation in professional skill (OR 1.63, 95% CI 1.12-2.4), have more chance of leaving early. General Culture (OR 1.70, 95% CI 1.20-2.39) is also a strong indicator because those who have higher evaluation show 70% more chances of prematurely leaving.

There is still the need to develop a linear mixed model to accommodate the longitudinal nature of the covariates that describe the periodic evalua- tions.

Learning X given Y, in the absence of Training Data

Cedric Spire, Dalia Chakrabarty
University of Leicester

Many data problems are driven by the will to find predictive relationships between predictors and response variables, using available information, where the aim is to predict the value of the response variable. In the paradigm of inverse problems, such is accomplished by learning this functional relationship between ***X*** and ***Y*** and thereafter computing the inverse of this function at the data. In many cases, training data is available, i.e. a set of values of ***X*** and the corresponding ***Y*** is at hand, allowing the supervised learning of the functional relationship based on that training data. However as is the case for physical systems for which only measurements on the observable ***Y*** are known, training data are not always available, rendering the task of modelling the predictor-response relationship, more challenging. Given that information on the link between ***X*** and ***Y*** is absent in the available data, we need to invoke this link from elsewhere; such is possible for systems of particles that abide by kinetic equations.

In this presentation, we will focus on a new unsupervised Bayesian learning method based on the modelling of the evolution of the state space density, in a situation for which training data are not available and only missing and noisy data is. Applications include the learning of dark matter density in real galaxies using data on astrophysical properties of galactic particles, learning of examinee ability distribution and item characteristics of a test, using real test scores, etc. The ulterior aim is to employ the learnt model parameter vector to the generation of training data - and thereafter, the supervised learning of the functional relation between observable and model parameter.

My poster however will be confined to a discussion of the method.

Multiple Imputations Technique in Missing Data Analysis: A Bayesian Approach Using Conjugate Prior

Athanasius Opara, Raymond Okafor, Ismail Adeleke
Distance Learning Institute, University of Lagos, Lagos, Nigeria

Rubin's (1987) multiple imputation (MI) is a popular incomplete data analysis method owing to its relative ease of application. Considerable research has been done by various authors to extend MI beyond Rubin's framework in order to model various data structures. This study contributes to the effort by extending MI to utilise available conjugate prior information on model parameters. The paper considers MI for linear regression model, with missingness in the response vector Y. Rubin's multiple imputation of missing Y starts by estimating posterior distributions of model parameters and drawing values from these posteriors to determine the conditional distribution of the missing data. Drawings are then made from this conditional distribution to impute the missing components. Non-informative prior distribution which is a scaled inverse chi-squared distribution was used in the Bayesian aspect of Rubin's approach. The study used joint normal inverse gamma distribution to model conjugate prior information in MI. Both simulated and real life data showed that estimates of linear model parameters, based on multiple imputation with conjugate information on model parameters, have precision that is quite competitive.

Exploring variable importance in a priori selection of candidate variables for species distribution models (SDMs)

Ramethaa Pirathiban, Kristen Williams, Anthony Pettitt, Samantha Low Choy
Queensland University of Technology

Recent reviews of SDM techniques have sought to optimize predictive performance. However, the extent to which such models reflect real-world species distributions also depends heavily on the relevance and quality of the input data. This confounding effect on model performance is exacerbated by current rapid increases in the number of potential predictor variables available: through improved land modelling, remote imagery and acoustic sensing. Attention given to variable selection has potentially great impacts, for instance, when the model is applied for prediction beyond the scope of the training data or for explanation.

In the last decade, the challenge of variable selection in SDMs has increasingly been addressed through post-hoc model comparisons, via indicators such as AIC (Akaike Information Criterion), and has been strongly advocated from an information-theoretic perspective. Typically, SDM users will either consider a few ecologically justifiable or easily measured predictors with the resultant potential for under-specified models. Otherwise, they draw upon a more comprehensive set of variables for input into popular data-driven approaches, such as Boosted Regression Trees and MAXENT, with the risk of over-fitting. These automated procedures do not necessarily select the best set of explanatory variables. Rather, they are sensitive to the criteria used for examining the usually infeasible number of all possible combinations of variables.

Bayesian SDMs do exist, with several methods previously considered for eliciting and encoding priors on model parameters. However, few methods have been published for informative variable selection; one exception is Bayesian trees. Here we develop and refine implementation of an elicitation protocol for variable selection in SDMs that helps makes explicit *a priori* expert judgements on the ecological relevance and the quality of candidate variables. We demonstrate how this information can be obtained then define priors and contribute to posterior analysis within Bayesian SDMs.

Simultaneous Bayesian Box-Cox Quantile Regression

Aziz Aljuaid, John Paul Gosling, Charles C Taylor
University of Leeds

Quantile regression is a powerful statistical method used to investigate the full conditional distribution of a response variable, and, as such, it provides more information than ordinary least squares regression. Quantile regression offers a comprehensive image of conditional distribution of the response variable by describing the relationship between the response variable and covariates at different quantile levels. Also, it is more robust to outlier and extreme observations. Therefore, it has been applied in a variety of fields such as science, finance, econometrics and environmental science.

In the context of Bayesian quantile regression based on the asymmetric Laplace likelihood, it is assumed that the relationship between the response variable and the covariates is linear with homoscedastic errors. To deal with violations of these assumptions, we use a Box-Cox transformation to develop a Bayesian method based on a pseudo-asymmetric Laplace likelihood to fit multiple quantile regressions. The issue of crossing quantile curves is investigated, and a solution based on prior constraints is considered. The proposed methods show an excellent capacity to handle the two violations of Bayesian linear quantile regression assumptions, which are non-linearity and heteroscedasticity. They can provide a smooth and reliable estimation of quantile functions, for complex non-linear models, without crossing over unbounded space of covariates. The key features of the proposed methods are their easy implementation and flexibility in the underlying assumptions.

Data and Methods Services: supporting research and providing training

Vanessa Higgins, Jo Wathan
UK Data Service

This poster highlights a number of ESRC-funded data and methods services which support research and study. These include:

- the UK Data Service

- Understanding Society

- CLOSER

- The Administrative Data Research Network

- Census and Administrative data LongitudinaL Studies Hub

- The National Centre for Research Methods

- The Centre for Longitudinal Studies

- Phase 2 of the Big Data Network.

These are valuable services that provide the resources and training needed to access high quality socio-economic data and to develop research skills.

Modelling Growth of Biological Processes Using Alternative Nonlinear Growth Models

Oluwafemi Oyamakin, Angela Chukwu
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 of the form; dH/dt = H[r + θ/√1+t^2]. 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.

A proposal for the multivariate evaluation of academic performance based on the analysis of publications of the last 30 years

Greibin Villegas Barahona, Galindo Purificación
Universidad de Salamanca

Retention and completion of university studies could be an overall indicator of quality of education in a country, and a direct way to assess the performance of economic investment in a country or region. The early abandonment of university studies causes frustration and demotivation; This indicator is a tool to assess the academic performance can have a population of students.

It is necessary to reconsider the concept of quality of education, because although it tends to increase the number of students in universities every year, it is certain that the increase in dropout rates and unsurpassed subjects also increases. One way to assess the issue of quality of higher education is assessing the academic performance of students with their study as possible an approximation to the educational reality.

The research has been made in this regard in the past 30 years, have similar characteristics regarding their conceptual, methodological and many cases similar results in building up the significance of some of the variables under study. A general conceptual criticism made to the design of most research is that researchers focus on a set of variables, conduct their research with a statistical technique and present the results in search of a statistical model that supports them make decisions.

In this sense, the research presented at this time to explain the academic performance of the student to be effective must have a construction platform data that supports a model of academic management in which action decisions that really impact in can academic performance product adjustments to the relevant variables.

Latent class modelling of longitudinal data: the challenges

Hannah Lennon, Matthew Sperrin, Andrew G Renehan
University of Manchester

Latent class longitudinal models can be used to cluster BMI trajectories to give a simplified representation of complex longitudinal structure. For BMI measurements as an example, this could correspond to 4 clusters of weight stable individuals, steady weight gainers, fast weight gainers and steady losers. Latent class models allow us to model heterogeneity through latent classes of trajectories. Latent class methods derive model-based clusters, unlike other clustering algorithms.

Latent class models are finite mixture models and they can be specified with and without random effects. The latent class mixed model assumes that the population is heterogeneous and composed of latent classes of subjects characterised by different mean profiles of trajectories. The model is an extension of the linear mixed model of Laird & Ware (1982) where both the fixed effects and the distribution of the random effects can be class-specific.

Different assumptions of the random effects distribution can be made and the assumptions are difficult to verify. Therefore specification is challenging. Many different variance structures can be assumed varying from unstructured, to constraining classes to allow class-specific intensities of variability. In some scenarios, constraints are necessary to aid convergence or for parsimony. However, the resulting latent classes that are derived depend much on these underlying assumptions.

We provide an illustration of the challenges of modelling the random structure correctly with the AARP cohort. We highlight that different assumptions can lead to very different results and hence inferences. In particular, we consider a simulation study to understand the effects of misspecification the random structure.

In conclusion, the distributional assumptions on the random effects should be carefully considered, while the interpretation of the latent classes should be kept simple and care taken to avoid over interpretation.

Modelling post-operative risk - when does the excess hazard return to baseline?

William Hulme, Glen Martin, Matthew Sperrin
Health eResearch Centre, University of Manchester

Patients undergoing interventional medical procedures are exposed to increased morbidity and mortality risks during the procedure and in recovery. This increase is tolerated because, on average over the longer-term, it is understood that the procedure improves patient outcomes. Quantifying the excess risk associated with the procedure is informative; knowing its magnitude and duration, and how it changes across different patient groups, operative practices, and indications, can be used to inform best-practise.

Relative Survival analysis offers a methodological framework to model this excess risk. Broadly, this can be achieved by comparing, via the hazard rate ratio, observed mortality with expected mortality as derived from patient-matched life table data. However, relative survival methods are typically used to study survival differences following disease diagnosis or to compare to long-term treatment strategies, where underlying hazards change slowly. These methods are not compatible with hazards arising from procedures carrying considerable short-term risks, as the initially high and rapidly decreasing mortality rate cannot be adequately modelled.

This work explores some modifications to standard relative survival approaches that are able to account for highly-skewed hazard rates typical of survival distributions arising from interventional procedures. These approaches are used to model the excess post-procedural risk associated with percutaneous coronary internventions (PCI), a common procedure in both elective and emergency settings to improve arterial patency in people with coronary artery disease. PCI data are taken from the British Cardiovascular Intervention Society PCI registry, and expected (baseline) hazards are derived from life table data published by the Office for National Statistics, matching on age, sex and year.

Spatial Modelling of Fever Prevalence and Suspected Malaria Cases among Children in SNNP and Oromia Regional States, Ethiopia

Aklilu Toma Shamenna
Hawassa University

Background: Disease morbidity, mortality and speed of spread vary substantially spatially. These have important implications for effective planning and targeting intervention strategies. The purpose of this study was to model the spatial dependence of fever prevalence and suspected malaria cases among children in Ethiopia.

Method: Data were obtained from 2011 EDHS collected for 144 districts at SNNP and Oromia Regional States. Explanatory spatial data analysis and spatial lag and error models were applied.

Results: The results showed that the spatial lag model better fitted to the data. Prevalence rate of each of the events in a district was shown to be affected by that of its neighbor's status. It was revealed that altitude, access to piped water, proportion of children under five, vaccination coverage, child wasting score, proportion of children born below average size and toilet availability were significant risk factors of fever rate. Moreover, altitude, proportion of children born below average, vaccination coverage, stunting score, wasting score, proportion of children under five, mother education, and access to mass media were found to have significant effects on the rate of suspected malaria cases.

Conclusions: There is spatial dependency for both variables -childhood fever prevalence and suspected malaria cases. The hot spot areas are at the center of each region. Several risk factors need attention. Interventions to mitigate occurrence of malaria infection among children would take in to account the nature of spatial variability and the identified risk factors.

Bayesian Network Analysis in Behavioural Clinical Veterinary Studies

Jf Collin
CEVA

The group Ceva Animal Health is one of the leaders in the treatment of animal behaviour issues such as unwanted cat scratching, dog travelling or urine marking. In the last recent years, the R&D of the group has explored an innovative path to address this field of research with the use of pheromone compounds. This topic is relatively new in the veterinary pharmaceutical industry and there is no recommended guideline proposed by the health authorities to conduct a clinical study with behaviour targets as primary outcome. Since animal behaviours are influenced by a high number of variables, clinical studies need to take into account multiple behavioural and environmental parameters that might be related. Those studies are conducted with the help of questionnaires to specific owners of animal(s) that record hundreds of variables during the trial. The aim is to determine the efficacy of a new product but also to discover the (conditional) relations between the environment and the different behaviours of the animals. As the potentially conditional relationships grow exponentially with the number of variables, we use Bayesian Networks as efficient unsupervised algorithms to explore those large multiparametric domains. This poster presents those methods and the results that we obtained with such analysis on a recent clinical trial. In particular we show how we validated our questionnaires, understand the relationships between different animal behaviours and asses the efficacy of a product with this statistical methodology.

Inhibitory geostatistical designs for spatial prediction taking account of uncertain covariance structure

Michael Chipeta, Dianne J. Terlouw, Kamija S. Phiri, Peter J. Diggle
Lancaster University

The problem of choosing a spatial sampling design for investigating unobserved spatial phenomenon S arises in many contexts, for example in identifying households to select for a prevalence survey in order to study disease burden and heterogeneity in a study region D. We study randomised inhibitory spatial sampling designs to address the problem of spatial prediction whilst taking account of the need to estimate covariance structure. Two specific classes of design are *inhibitory designs* and *inhibitory designs plus close pairs*. In an inhibitory design, any pair of sample locations must be separated by at least an inhibition distance δ. In an inhibitory plus close pairs design, $n - k$ sample locations in an inhibitory design with inhibition distance δ are augmented by k locations each positioned close to one of the randomly selected $n - k$ locations in the inhibitory design, uniformly distributed within a disc of radius ζ. We present simulation results for the Matérn class of covariance structures. When nugget variance is non-negligible, inhibitory plus close pairs designs demonstrate improved predictive efficiency over designs without close pairs. We illustrate how these findings can be applied to the design of a rolling Malaria Indicator Survey that forms part of an ongoing large-scale, five-year malaria transmission reduction project in Malawi.

Effects of Some Design Factors on the Distribution of Similarity Indices in Cluster Analysis

Ahmed Albatineh, Hafiz Khan, Bashar Zogheib, Golam B.M. Kibria
Kuwait University

This paper investigates the effects of number of clusters, cluster size, and correction for chance agreement on the distribution of two similarity indices, namely, Jaccard (1912) and Rand (1971) indices. Skewness and kurtosis are calculated for the two indices, their corrected forms and compared with those of the normal distribution. Three clustering algorithms are implemented: complete linkage, Ward, and K-means methods. Data were randomly generated from bivariate normal distribution with specified mean and variance covariance matrix. Threeway ANOVA is performed to assess significance of the design factors using skewness and kurtosis of the indices as responses. Test statistics for testing skewness and kurtosis and observed power are calculated. Simulation results showed that independent of the clustering algorithms or the similarity indices used, the interaction effect cluster size×number of clusters and the main effects of cluster size and number of clusters were found always significant for skewness and kurtosis. The three way interaction of cluster size×correction× number of clusters was significant for skewness of Rand and Jaccard indices using all clustering algorithms, but was not significant using Ward's method for both Rand and Jaccard indices, while significant for Jaccard only using complete linkage and K-means algorithms. The correction for chance agreement was significant for skewness and kurtosis using Rand and Jaccard indices when complete linkage method is used. Hence, such design factors must be taken into consideration when studying distribution of such indices.

Inference on covariance operators via concentration inequalities

Adam Kashlak, John A. D. Aston, Richard Nickl
University of Cambridge

We propose a novel approach to the analysis of covariance operators making use of concentration inequalities. First, non-asymptotic confidence sets are constructed for such operators. Then, subsequent applications including a k-sample test for equality of covariance, a functional data classifier, and an expectation-maximization style clustering algorithm are derived and tested on both simulated and phoneme data.

Accounting for measurement error in biomarker data and misclassification of subtypes in the analysis of tumor data

Daniel Nevo, David Zucker, Molin Wang, Rulla Tamimi
Harvard T.H. Chan School of Public Health

A common paradigm in dealing with heterogeneity across tumors in cancer analysis is to cluster the tumors into subtypes using marker data on the tumor, and then to analyze each of the clusters separately. A more specific target is to investigate the association between risk factors and specific subtypes and to use the results for personalized preventive treatment. This task is usually carried out in two steps – clustering and risk factor assessment. However, two sources of measurement error arise in these problems. The first is the measurement error in the biomarker values. The second is the misclassification error when assigning observations to clusters. We consider the case with a specified set of relevant markers and propose a unified single-likelihood approach for normally distributed biomarkers. As an alternative, we consider a two-step procedure with the tumor type misclassification error taken into account in the second-step risk factor analysis. We describe our method for binary data and also for survival analysis data using a modified version of the Cox model. We present asymptotic theory for the proposed estimators. Simulation results indicate that our methods significantly lower the bias with a small price being paid in terms of variance. We present an analysis of breast cancer data from the Nurses' Health Study to demonstrate the utility of our method.

Planning for the future – Census data, Population data, Geographical data

Jo Wathan
University of Manchester

Census enumeration is starting to change from the monolith that was traditional census-taking, to the gathering data from a wider range of sources. It is important that at this time of change, we have a clear understanding of research users' data needs.

The Census Support service (part of the UK Data Service) and its precursors have always provided access to related data, but we are keen to ensure that this work is expanded to maintain and improve data utility in a period of census transformation.

This poster will cover the following

- Services provided by Census Support
- Work underway to improve the flexibility of our tools, to enable users to query multiple data types and to generate linked outputs
- Identify datasets already supported by Census Support that are from sources other than the census
- Current priorities for integrating additional data sources into our systems to improve ease of use
- Information on our work to establish user needs going forward, and how you can make your views known.

Is ANOVA robust to extreme departures from normality?

Jaume Arnau, Rafael Alarcón, Roser Bono, Rebecca Bendayan, María J. Blanca
University of Barcelona

The robustness of analysis of variance to non-normality has been studied since the 1930s through to the present day. However, this extensive body of research has yielded contradictory results, there being both evidence for and against the robustness of *F*-test to deviations from normality. More specifically, the evidence against suggests that *F*-test is only robust with moderate departures from normality and provided that the populations have equal distributions. The aim of this study was to determine the effect on Type I error of several extreme deviations from normality, considering both equal and unequal group distributions. To this end, we conducted a Monte Carlo simulation study involving a design with three groups and several known distributions. The manipulated variables were: balanced and unbalanced design; group sample size and total sample size; coefficient of sample size variation; type of distribution (exponential, double exponential and chi-square with 8 degrees of freedom); and equal and unequal distribution in the groups. Data were generated using a series of macros created ad hoc in SAS 9.4. Ten thousand replications of each combination of the manipulated conditions were performed at a significance level of .05, recording the empirical Type I error rate. Bradley's liberal criterion was used to assess the robustness of the procedure. The results showed that *F*-test was robust in 100% of the cases studied, independently of the type of distribution, equal or unequal distributions, equal or unequal group sample sizes or the coefficient of sample size variation. In conclusion, *F*-test is not sensitive to departures from the normality assumption with extremely contaminated distributions such as those considered in this study.

This research was supported by grant PSI2012-32662 from the Spanish Ministry of Economy and Competitiveness.

Generalized linear mixed models in longitudinal studies (1996-2015): Systematic review and bibliometric analysis

Roser Bono, María J. Blanca, Jaume Arnau, Rebecca Bendayan, Rafael Alarcón, Dolores López-Montiel
University of Barcelona

Generalized linear mixed models (GLMMs) are useful for analysing correlated data, including longitudinal data and repeated measures. These techniques do not require the error terms to be normally distributed, and they are well-suited to most of the distributions found with quantitative and categorical data in applied research. With the aim of seeing how GLMMs have been used over time in various fields we present a systematic review of their application in repeated measures designs. The search was carried out in the Web of Science (WOS) database for the period 1996-2015, without restrictions on language or publication type. The review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) checklist. Documents were selected on the basis of the title and the abstract by two independent reviewers (kappa coefficient = .80). A total of 199 studies were included and were categorized according to year of publication, name of the journal, type of study (theoretical, empirical or simulation) and field of application. The results showed a notable growth in the number of theoretical publications from 2004 onwards, with empirical studies, as well as simulation studies analysing robustness and power, beginning to be published since 2009. The largest proportion of studies corresponded to the statistical field, and it is only in recent years that studies related to the health and social sciences have begun to be published. This is corroborated by the distribution of scientific journals (Bradford's law), since 34% of articles in the first zone were published in statistical journals, the three most common being *Biometrics*, *Computational Statistics and Data Analysis* and *Statistics in Medicine*.

This research was supported by grant PSI2012-32662 from the Spanish Ministry of Economy and Competitiveness.

Bayesian Bounds for Population Proportion under Ranked Set Sampling

Vivek Verma, Dilip C. Nath, Radhakanta Das
Department of Statistics, Gauhati University, India

This paper presents a framework for Bayesian estimation of population proportion p, and a novel method for its comparison with the classical estimate using ranked set samples. The attempt of this paper is to derive the Cramer Rao bound for population proportion, p, where p may be considered as a fixed or random component. Based on the conditions that p is a non-random or random parameter, maximum likelihood estimator (MLE) and maximum a posteriori estimator (MAP) are obtained, respectively. The performance of the corresponding ML and MAP estimator are evaluated by comparing the bounds of variance of estimator which is obtained through Cramer Rao (CR) bound and Bayesian Cramer Rao (BCR) bound of the estimator, respectively. Simulation results indicate that the derived BCR bound provides a compact lower bound for the estimator of p. Also the application of the obtained bounds is illustrated using a real-life data in relation to estimate the non-immunization of measles vaccine among children in the rural India.

Statistical Machine Learning Approach to Correlated Time Series

Omar Alzeley
University of Leicester

Minimizing a constrained multivariate function is the fundamental of Machine learning, and these algorithms are at the core of data mining and data visualization techniques. The decision function that maps input points to output points is based on the result of the optimization. This optimization is the central of learning theory. The analysis of tree rings has been used in diverse fields such as the dating of historical buildings and environmental studies. It is defined as the science based on the fact that every growing season a tree adds a new layer of wood to beneath its bark. Over a period of several successive growing periods, usually years, the sequence of tree ring widths form a time series. The purpose of this research is to analyses a set of tree ring widths data from nine trees growing in Nottinghamshire and discuss how such time series can be modelled by using machine learning algorithms.

Firstly, we have used descriptive statistics and visualization to understand the data. We used data transforms in order to better expose the structure of the prediction problem to modeling algorithms. The algorithms we used have been evaluated. Secondly, we used algorithm tuning and ensemble methods to improve our results. Finally, we compared the classes method, which has been used on the same dataset in previous study, with the methods we used in this research.

Under 5 Mortality Rates in G7 countries: Analysis of Fractional Persistence, Structural Breaks and Non-linear Time Trends

OlaOluwa Simon Yaya, Luis Gil-Alana, Yaw Amoateng
University of Ibadan, Nigeria & North West University (Mafikeng Campus), South Africa

This paper deals with the analysis of the under 5 mortality rate series in the G7 countries by means of using fractional integration techniques, including structural breaks and potential non-linearities in the data. Several features were detected throughout the results: firstly we observed that for the neonatal data, the order of integration is equal to or higher than one in all cases, contrary to what happens for the remaining cases (<1 to <5 years) where mean reversion is found in many cases, especially as we increase the age of death. Thus, shocks affecting the neonatal (<1 month from delivery) mortality rates will have permanent effects requiring special attention to recover the original trends. As expected, all the time trend coefficients were significantly negative and the highest reduction in the mortality rates was obtained in the case of Japan, which is as a result of a 17-year increase of life expectancy for the country. The implications of the findings for international, national and local agencies studying population dynamics, particularly those with intervention programmes to neonates, and also to premium insurers who rely on life expectancy, are discussed. Due to the sensitivity of the methodological approach applied in this study, it is recommended that a robust time series approach should always be employed for analyzing child mortality rates.

Efficient High-Dimensional Drug Sensitivity Prediction in Heterogeneous Cancer Cell Lines

Frank Dondelinger, Sach Mukherjee
Lancaster University

Recent advances in high-throughput drug screening have led to increasing interest in drug sensitivity prediction from cell line molecular data (Barretina et al., 2012; Garnett et al., 2012; Costello et al., 2014). The benefits of computational statistics and machine learning approaches are apparent: the ability to predict measurements on unseen cell lines precludes expensive experiments; studying the structure of the models can reveal novel biomarkers for precision medicine; and analysing the performance of drug-specific models on different tumor types will indicate potential new targets for drugs. The challenges of drug sensitivity prediction reside in the inherent heterogeneity of the data, both across and within datasets.

Heterogeneity across datasets can often be traced back to different technologies and measurement protocols (Hatzis et al., 2014). This complicates the task of combining multiple drug sensitivity datasets to obtain greater predictive power. Heterogeneity within datasets arises due to the existence of biological differences between the samples (cell lines) in a dataset; for instance when combining different tumor types in one dataset.

We present a principled method for modelling prediction in heterogeneous settings, based on information sharing across penalised linear regression models. This allows us to build prediction models that reflect the inherent heterogeneity in the data, while at the same time leveraging the commonalities between groups of homogeneous samples. We present two related approaches, using l1 and l2 fusion penalties on the model-specific parameters, and show in extensive simulation studies that these approaches outperform both naive pooling and complete separation of models in realistic scenarios. We then apply our method to gene expression and drug sensitivity data from The Cancer Cell Line Encyclopedia (Barretina et al., 2012), where the different models represent different tissue types, and we demonstrate that out method improves on other popular prediction approaches, as well as being highly computationally efficient.

Algorithmic Design for Big Data: The ScaLE Algorithm

Murray Pollock, Paul Fearnhead, Adam Johansen, Gareth Roberts
University of Warwick

This poster will introduce a new methodology for exploring posterior distributions by modifying methodology for exactly (without error) simulating diffusion sample paths. This new method has remarkably good scalability properties as the size of the data set increases (it has sub-linear cost, and potentially no cost), and therefore is a natural candidate for “Big Data” inference.

Joint work with Paul Fearnhead, Adam Johansen and Gareth Roberts.

Modelling enterprise level statistics using administrative data

Megan Pope, Jonathan Digby-North, Gary Brown
Office for National Statistics

The UK is required to provide estimates of Structural Business Statistics (SBS) to Eurostat on an enterprise reporting basis for the 2016 reference year. This represents a new output, as currently the Office for National Statistics (ONS) produces estimates on a Reporting Unit (RU) basis. For those enterprises with more than one RU, any intra-flows within the enterprise must be removed to produce consolidated accounts. A feasibility study was conducted to investigate whether it would be viable to estimate enterprise level statistics (e.g. turnover) using existing data.

Two external/administrative sources were identified as able to provide enterprise level data: FAME and VAT. FAME is produced by Bureau Van Dijk and VAT data is provided to ONS by HMRC. The data were matched to the Inter-Departmental Business Register (IDBR) and multiple regression modelling was used to derive relationships between RU level ABS/IDBR variables and the external enterprise level data.

In general, total RU turnover and employment were strong predictors of consolidated turnover. When applying the models to ABS data, consolidated turnover was found to be approximately 10-20% lower than the simple sum of RU turnover. Although the findings have been more or less consistent across data sources, there are data quality and implementation issues which require further work to resolve.

Bayesian modelling and computation for surface-enhanced Raman spectroscopy

Matt Moores, Kirsten Gracie, Jake Carson, Karen Faulds, Duncan Graham, Mark Girolami
University of Warwick

Raman spectroscopy is a technique for detecting and identifying molecules such as DNA. It is sensitive at very low concentrations and can accurately quantify the amount of a given molecule in a sample. The pattern of peaks in a Raman spectrum corresponds to the vibrational modes of the molecule. The shift in frequency of the photons is proportional to the change in energy state, which is reflected in the locations of the peaks. Each Raman-active molecule has a unique spectral signature, comprised by the locations and amplitudes of the peaks. The presence of a large, nonuniform background presents a major challenge to analysis of these spectra. The amplitudes of the peaks are completely dependent on the position of the baseline and vice-versa.

We introduce a sequential Monte Carlo (SMC) algorithm to separate the observed spectrum into a series of peaks plus a smoothly-varying baseline, corrupted by additive white noise. The peaks are modelled as Lorentzian or Gaussian functions, while the baseline is estimated using a penalised cubic spline. Our model-based approach accounts for differences in resolution and experimental conditions. We incorporate prior information to improve identifiability and regularise the solution. The posterior distribution can be incrementally updated as more data becomes available, resulting in a scalable algorithm that is robust to local maxima. These methods have been implemented as an R package, using RcppEigen and OpenMP.

Our model has enabled us to directly estimate quantities of scientific interest. By incorporating this representation into a Bayesian functional regression, we can quantify the relationship between molecular concentration and peak intensity. We also calculate the model evidence using SMC to investigate long-range dependence between peaks. The Bayes factor indicates that the Lorentzian function is better supported by the observed data, which highlights the need for joint models to quantify Raman spectroscopy.

Rubrics and in-class feedback: students' perceived usefulness for learning in a research designs course

Roser Bono, M. Isabel Núñez-Peña, Macarena Suárez-Pellicioni
University of Barcelona

Recent research has highlighted the importance of feedback for improving learning. This study examines students' views regarding two types of feedback: that obtained through rubrics and that given by the class tutor (rubrics and in-class feedback, respectively). With this aim we constructed an ad hoc questionnaire to assess the perceived usefulness of both types of feedback in relation to the following aspects: a) course preparation, b) improved learning, c) awareness of assessment criteria, d) self-assessment, e) reduced exam anxiety, and f) need for their use in other courses. The sample comprised 135 undergraduates from the University of Barcelona who were enrolled in a Research Designs course as part of the degree in Psychology. Of these, 106 stated that they had used the rubrics, 129 had attended feedback classes and 100 had both consulted the rubrics and attended feedback classes. On almost all the questionnaire items the percentage of positive views was high for both types of feedback, and especially for in-class feedback. The only item for which responses indicated a lower level of perceived usefulness was the one referring to reduced exam anxiety. Finally, we applied the Wilcoxon signed-rank test for related samples to the responses of the 100 students who had both used the rubrics and attended feedback classes. This analysis showed that the feedback class was viewed more positively than were rubrics for all the analysed aspects (p<.02), with the exception of the item about knowledge of assessment criteria (p=.833).

This research was supported through projects 2014PID-UB/069 and GIDCUB-13/099 of the Consolidated Group for Innovation in Teaching at the University of Barcelona.

Modelling spatial heteroskedasticity by volatility modulated moving averages

Michele Nguyen, Almut Veraart
Imperial College London

Spatial heteroskedasticity, which refers to changing variances and covariances in space, is a feature that has been observed in environmental data. While prominent models in the literature have accounted for this behaviour by multiplying a Gaussian error process with a stochastic volatility process, we propose a model that intricately blends the effects of spatial volatility across space. This has a more natural link to stochastic partial differential equations.

Let $\mathbf{t} \in \mathfrak{R}^d$ for some natural number d. Our model, which we call the volatility modulated moving average (VMMA), is defined by:

$$Y(\mathbf{t}) = \int_{\mathfrak{R}^d} g(\mathbf{t}-\mathbf{s})\, \sigma(\mathbf{s})\, W(d\mathbf{s}),$$

where g is a deterministic (kernel) function, W is white noise and σ is a stationary stochastic volatility field, independent of W. Without σ, this model reverts to the Gaussian moving average which is frequently used in Geostatistics to design covariance structures.

In this project, we develop a discrete convolution simulation algorithm for the VMMA as well as a two-step moments-based inference procedure based on its theoretical properties. As an illustration, we apply our method to sea surface temperature anomaly data.

Predicting daily carbon emission from electricity generation in the UK using functional data analysis and state space models

Zehui Jin
University of Manchester

The data used to compute the carbon emission from the electricity generation in the UK are the electricity geneartion by fuel in every half hour. The daily generation by fuel is a vector of 48 dimensions. It is natural to think of the daily generation as a continuous process, so the vector-valued data is transformed into the functional data by smoothing. The variation of the function of the daily generation from day to day and a prediction of the functions for the next few days is the object of interest. To achieve that, functinal principal component analysis is employed such that the variation of the functions can be learned by their principal component scores. State space models are used to model the evolution of the principal component scores along time and give predictions. State space models are shown to be more flexible and more precise in prediction than ARIMA models in this study.

Global burden of drug-resistant tuberculosis in children

Pete Dodd, Charalambos Sismanidis, James Seddon
University of Sheffield

Background After infection with *Mycobacterium tuberculosis*, children are at increased risk of progression to tuberculosis disease, which can be challenging to diagnose. New estimation approaches for children have highlighted the gap between incidence and notifications of *M tuberculosis*, and suggest more cases of isoniazid-resistant and multidrug-resistant (MDR) disease than are identified. No work has yet quantified the burden of drug-resistant infection, accounted for other types of drug resistance or sampling uncertainty.

Methods We combined a mathematical model of tuberculosis in children with an analysis of drug-resistance patterns to produce country-level, regional, and global estimates of drug-resistant infection and disease. We determined drug resistance using data from the Global Project on Antituberculosis Drug Resistance Surveillance at WHO between 1988 and 2014. We combined 1000 sampled proportions for each country from a Bayesian approach, with 10,000 sampled country estimates of tuberculosis disease incidence and *M tuberculosis* infection prevalence. We estimated the proportions of tuberculosis cases at a country-level with isoniazid monoresistance, rifampicin monoresistance, multidrug resistance (MDR), fluoroquinolone-resistant multidrug resistance, second- line injectable-resistant multidrug resistance, and extensive multidrug resistance with resistance to both a fluoroquinolone and a second-line injectable (XDR).

Findings We estimated 850,000 children developed tuberculosis in 2014; 58,000 with isoniazid-monoresistant- tuberculosis, 25,000 with MDR-tuberculosis, and 1200 with XDR-tuberculosis. We estimate 67 million children are infected with M tuberculosis; 5 million with isoniazid monoresistance, 2 million with MDR, and 100,000 with XDR. Africa and southeast Asia have the highest numbers of children with tuberculosis, but the WHO Eastern Mediterranean, European, and Western Pacific regions also contribute substantially to the burden of drug-resistant tuberculosis because of much higher resistance proportions.

Interpretation Far more drug-resistant tuberculosis occurs in children than is diagnosed, and there is a large pool of drug-resistant infection. This finding has implications for empirical treatment and preventive therapy in some regions.

The Constituency Explorer

Jim Ridgway, Richard Cracknell, Sinclair Sutherland, James Nicholson
Durham University

The Constituency Kit arose from a collaboration between Durham University and the House of Commons Library. We created resources that provide rich data in a form that people can understand and explore, ahead of the 2015 election. The Constituency Kit targets resources at different audiences, that include politicians, journalists, researchers, and citizens. A key ambition was to create resources that people actually use. The Constituency Kit http://www.constituencyexplorer.org.uk/ has three major components:

The Constituency Explorer - a dynamic visualisation that presents information on the location of every constituency (arranged in NUTS regions) on 150+ variables. Users can search by name or via an interactive map, and choose variables via a pull-down menu. Metadata are immediately available; data can be downloaded; there are links to social media; users can 'pin' constituencies to make comparisons. Users can embed the Explorer in their own websites.

Data are disaggregated by constituency, and users can now (post election) toggle between the 2010 and 2015 election results.

Quizzes - run on smartphones and ask 7 questions about a constituency. Results can be posted to social media. The quiz was designed to entice everyone to think politically, via a game.

Constituency PDFs - each of these is a downloadable House of Commons Library Report (with a total of around 10,000 pages).

The Explorer has been used extensively (e.g. 1000 users on day 1, accessing over 3000 pages) and has been widey cited, and has been used in teaching in a broad range of disciplines.

The Poster will map out the features of the Constituency Kit, and visitors will be encouraged to explore the resource for themselves.

Mean and variance of round off error

Rui Li, Saralees Nadarajah
University of Manchester

Gadzhiev derived expressions for round off error mean and round off error variance when the rounded variable follows the centered uniform and centered Gaussian distributions. Here, we derive general expressions for round off error mean and round off error variance when the rounded variable is any continuous random variable on the real line or any continuous random variable over a finite interval. Numerical studies are given.

Bayesian semi-parametric determination of the source of microbiological contamination in food pathways

Poppy Miller, Chris Jewell, Jonathan Marshall, Nigel French
Lancaster University

Numerous zoonotic diseases cause morbidity, mortality and productivity losses in both humans and animal populations. For many zoonotic diseases that are important to human health (such as campylobacteriosis), it is difficult to attribute human cases to sources because there is little epidemiological information on the cases. Genotyping systems allow the zoonotic pathogens to be categorised, and the relative distribution of the genotypes among the sources (food sources or reservoirs of bacteria) and in human cases allows inference on the likely source of each genotype. Current source attribution models, specifically the Island model (Wilson et al. 2008), Hald (Hald et al. 2004) and modified Hald models (Mullner et al. 2009) are not fully joint and have many (often unverifiable) assumptions. Identifiability of the parameters in this model is an issue because a large number of parameters need to be estimated, the data is imbalanced, and many of the combinations of source and type and have very low counts. We present techniques to overcome these issues within a Bayesian framework by developing a fully joint model which non-parametrically clusters the type effects (using a Dirichlet Process) allowing identification of groups of bacterial subtypes with similar pathogenicity, survival and/ or virulence mechanisms. This model is applied to Campylobacter data from the Manawatu area of New Zealand (previously analysed by Mullner et al. (2009)) using the sourceR package, and compared to current source attribution models.

Analysing Bird Traps and Bird Behavioural Data

Zhou Fang
Biomathematics and Statistics Scotland

Certain types of bird traps are used as a pest control measure by farmers, gamekeepers and others, targeting corvids such as crows (Corvus corone) and Magpies (Pica pica). We investigated the performance of traps on arable farmland. Firstly, to compare the ability of the traps at catching birds, and secondly, to ensure that the traps were humane and did not have excessively negative welfare implications.

To do this, an experiment was conducted where three types of trap were placed at different sites over the course of a year, trapped birds were monitored by camera to observe behaviour and were assessed physically when removed from the traps.

We undertook analysis based on generalised linear models to compare trap catch rates relative to the prevalence of target birds. We recorded the behavioural state of the birds, at set time intervals and counted the numbers of actions taken of each type during regular samples. We constructed graphs to explore the behaviour data, and modelled them so as to compare statistically differing behaviours between traps, and at different times of day. Further analysis involved confirming whether behavioural activity was linked the prevalence of injuries in the birds.

Our investigations suggested significant and important differences between the trap types employed, as well as other interesting aspects of trapped bird behaviour, that have implications for the future employment of traps.

Multidimensional smoothing and unknown changes in the pattern of response

Mario Martinez-Araya, Jinxin Pan
The University of Manchester

In longitudinal and spatial data analysis it is common that the trend of the response variable of interest, for instance Y, does not exhibit a linear pattern of relation with the explanatory variable or covariate, denoted X. This non-linear trend might be due to a truly non-linear relationship between Y and X, but also due to the existence of unknown changes in the pattern of response in the range of the covariate. For example in one dimension (1D) due to change-points or in 2D due to contours of change. In 1D, let g(X) be the observed change in the response Y at X. Then tentative models are: (a) the pure linear where g(X) is a straight line, (b) the segmented or linear with G unknown change-points, (c) the spline where g(X) is a unknown smooth-function, and (d) the spline with G unknown change-points. We propose a new mixed-model representation of splines which allows to model linear or segmented linear trends jointly with splines by including fixed-effects, random-effects and variance components. The smoothing parameter is the ratio of the variance components and is estimated using traditional maximum likelihood (ML/REML) instead of cross-validation approaches. These mixed-model representation are valid for any arbitrary dependence structure (i.e. ARMA(p,q) or Toeplitz) or variogram function (i.e. Matern, Powered-Exponential or Spherical) in the errors. To select among the models we propose exact tests for linearity (i.e. H0: pure linear or segmented linear against H1: spline or spline with change-points) and the existence of change-points (i.e. H0: G0 change-points against H1: G1 change-points, G0<G1). The method proposed overcomes limitations of existing methods which assume independent data or are only valid for a class of correlations, therefore not suitable for longitudinal and spatial data. Simulations, power evaluations and real data analysis with comparisons with current available methods are provided.

Spatially adaptive kernel density estimation: exploring global bandwidth selection techniques

Tilman Davies, Martin Hazelton, Claire Flynn
University of Otago, New Zealand

The highly heterogeneous nature of planar point patterns typically observed in many different fields of research renders nonparametric methods for estimation of the assumed bivariate density function particularly attractive. One often turns to kernel density estimation – whether it is simply for data exploration, or for the construction of key components of more complicated spatial models. Most often deployed with a constant amount of smoothing, recent advancements in KDE have highlighted clear advantages of allowing the amount of smoothing to vary spatially – where the adaptive bandwidths reduce to capture more detail in densely populated sub-regions, and increase in sparsely populated areas to 'smooth over' less important features. As with any smoothing problem, however, bandwidth selection is critical to the reliability of the results – and we cannot expect bandwidth selection methods developed for the fixed estimator to simply 'work out of the box' for choosing the requisite global smoothing parameter for the adaptive estimator. Sound data-driven solutions to this problem are essentially nonexistent at present and, backed up by some novel theoretical and numerical explorations, we provide some initial recommendations for implementation of the flexible adaptive estimator.

Assessing the use of an address based sampling design for the 2021 Census Coverage Survey

Adriana Castaldo, Owen Abbott, Ercilia Dini
Office for National Statistics

This poster describes the work that the Office for National Statistics (ONS) is doing to assess a proposed new sampling design for the 2021 Census Coverage Survey (CCS) of England and Wales. The CCS is a short interviewer-based doorstep survey of around 300,000 households conducted a few weeks after the census, designed to measure under-count of households and people in the census. The census and CCS data are used in a capture-recapture estimator to produce estimates of the total population by age, sex and Local Authority. The choice of sample design strategy for the CCS can have a major impact on the accuracy of the population estimates. One key innovation being considered for the 2021 CCS is the use of addresses rather than postcodes as the final sampling unit. The main advantage of using address based sampling for the CCS is better quality estimates due to smaller clustering effects. This was confirmed by preliminary results obtained through simulation studies. The main disadvantage is a potential increase in costs due to greater travel distances for the interviewers. Whilst the simulation results suggest that address based sampling is worth pursuing, the trade off between precision and cost requires further research. In addition, the wider impacts need to be understood which include, for example, the practicalities around its implementation.

Evidence of infectious transmission in inpatients infected with Clostridium difficile: what can incomplete data say?

Pietro Coen, Leila Hail, Gema Martinez-Garcia, Annette Jeanes
Infection Control office, UCLH NHS Foundation Trust

Objectives

Toxigenic *Clostridium difficile* causes morbidity (mainly diarrhea) and mortality (via gastrointestinal damage) in hospital inpatients. The organism survives in the environment as resistant spores and patients can be asymptomatic carriers. Assessment of the true extent of transmission in hospital is problematic because admission screening is not mandatory in the NHS (only symptomatic patients get tested); the environment is rarely tested. We used ribotyping and patient transition data to estimate transmission events in inpatients admitted to University College London Hospitals (UCLH) NHS Trust.

Methods

Between April 2014 to March 2016 the subset of first known positives within a 28 day window were classified as potential 'recipients' of infection; positive patients who overlapped in ward-stay with a 'recipient' were identified as 'donors'. Where donor-recipient pairs had the same ribotype we inferred potential transmission.

Results

For 711 of 823 recipients (86%) we identified potential 'donors' for a total of 4329 donor-recipient pairs. Some were directly exposed (529 who shared the same ward at the same time as the donor) and indirectly exposed (182 who did not overlap in ward stay but stayed on the same ward within 28 days of donor discharge – a proxy for environmental exposure). Only 281 directly exposed recipients had useful ribotyping data (53%): 28 of these had a ribotype match with a donor (10%) amounting to 6.4% of all recipients. Indirectly exposed recipients had 30 ribotype matches (28%) amounting to 6.1% of all recipients.

Conclusion

Ribotyping and patient stay data suggest that exposure to an infectious agent on the ward can account for up to 12.5% of all new cases of *C. difficile* cases in UCLH inpatients; the remaining 87.5% may have been "pseudo-acquisitions" (patients admitted as asymptomatic carriers) or may have acquired it from asymptomatic carriers.

Optimal detection

Peter Johnson
The University of Manchester

I will present an overview of optimal stopping in mathematical statistics. These problems include sequential hypothesis testing sequential change-point detection. The problems use stochastic calculus to provide sufficient statistics and optimal decision hitting boundaries but also prove that the solution is optimal in the sense that no other detection method will provide a quicker expected delay for a given probability of error (false alarm or missed detection probabilities can be bounded).

My research has mainly considered continuous time-processes (but which can tell us a lot about many discrete time processes as well) and I have been directly involved applying these problems to many areas; including detecting signals in underwater sonar systems, detecting radioactive material entering the UK (using Poisson processes) and detecting breakages in atomic clocks on satellites for the development of a European global positioning system (the Galileo project).

Detecting breakages in atomic clocks is a nice example of the importance of making decisions both as quickly and accurately as possible. If one of the clocks on board a GPS satellite gains an error of just 10 nano seconds than this relates to a 30m error on the ground. Detecting these error quickly and accurately so that the satellites can be re-calibrated helps mitigate risks on the ground and commercial costs of false positives. This is especially important as these systems are used to help land planes.

Using Validation, Imputation & Paradata to Inform Modal Effects when Moving to Electronic Data Collection on the Monthly Wages and Salaries Survey (MWSS)

Genevieve Hopkins, Iolo Tomlinson
Office for National Statistics

In line with the UK Government's 'Digital by Default' initiative, all ONS surveys and the 2021 Census will be moving online as the principle mode of data collection.

Currently, as with most ONS economic surveys, the MWSS is predominantly collected using paper questionnaires.

Between May 2015 and March 2016, a pilot study was carried out on the MWSS to assess the modal effects of moving to online data collection. In order to test for modal effects, a parallel run of the paper MWSS and an online version of the survey took place. From this a number of metrics were collected. This poster demonstrates the approaches taken to analyse modal effects and how this information was used to inform and improve the data collection process. The resulting findings will be presented in three areas:

Validation – Validation by mode by:

- reviewing error failure rates (including selective editing failures)
- analyzing re-contact 'hit' rates (data changed following re-contact)

Imputation – showing the developing methodology of various imputation techniques, including pattern imputation and mean imputation, across and within online and paper datasets, plus imputation rates by mode.

Paradata – demonstrating how paradata can be used for error analysis and to inform online questionnaire design. This was 'footprint' data collected giving information on the respondent journey through the online questionnaire. Factors assessed included triggering of soft and hard edit checks and time taken to complete.

Bayesian Calibration of Disease Models for Health Systems Evaluation

Daria Semochkina, Cathal Walsh
University of Limerick

One of the biggest objectives in modern life is to tracks patients as they are diagnosed each year and progress through the disease and treatment course. This project examines the calibration of natural history model of HPV virus progression using a fully Bayesian approach. The model's predictions of key outcomes depends on a number of unobserved parameters such as transition probabilities between health states. This model facilitates decision making regarding population level interventions such as screening and vaccination.

The calibration of a natural history model is an inverse problem where the parameters are estimated using data from observed cohorts of individuals at different stages of disease. Markov Chain Monte Carlo (MCMC) sampling was used to obtain samples from the distributions for these parameters. This was compared with a simple Monte Carlo approach. A number of different proposals within the MCMC framework were examined to find the best calibration method. Our motivation in this research is to use MCMC methods to explore the distribution of the transition rates, conditional on the information available from the cohort studies. Some discussion on the use of prior information regarding the physically plausible rates and non-identifiability isuues for some of the parameters is also discussed.

The MCMC approach is shown to be working better than the simple Monte Carlo independent sampling. However, within the MCMC framework different proposals showed different convergence and mixing properties. Finding the balance between changing proposals to achieve better results has proven to be a great challenge.

What census response rate is needed to produce high quality estimates?

Viktor Racinskij
Office for National Statistics

The key aim of the 2021 Census in England and Wales is to produce high quality population estimates by local authority and age-sex. The estimates are derived from a survey designed to measure coverage in the census, called the Census Coverage Survey (CCS). The estimates are then produced by matching the census and CCS data together in order to apply an estimation framework that combines dual system and ratio estimators.

This poster presents the research undertaken by the Office for National Statistics to explore the relationship between the census response pattern (mean response rate and the variance of the response) and the resulting precision of the population estimates. The intention is to use this research to help set response rate targets for the 2021 Census in order to provide assurance that the census population estimates will be of high quality.

Census response patterns were simulated from a multilevel logistic model fitted to the 2011 Census and CCS data. The model provided predicted probabilities of people and households responding in the census based on various characteristics. The simulations were designed to ensure that the mean and the variance of responses were controlled while preserving the original odds ratios structure of the modelled 2011 Census responses. Overall, 9,600 pairs of censuses and CCSs were simulated for a sample of areas and each pair fed into the estimation framework to produce population estimates for the chosen areas. These estimates were analysed to examine the relationship between the census response patterns and the precision of the estimates. Of particular interest was whether it was better to have less variability or a higher overall response rate.

This work provided statistical insights that will help to shape the response targets for the 2021 Census.

Bayesian Inference for Nonresponse Two-Phase Sampling

Nanhua Zhang, Yue Zhang
Cincinnati Children's Hospital Medical Center

Nonresponse is an important practical problem in epidemiological surveys and clinical trials. Common methods for dealing with missing data rely on untestable assumptions. In particular, non-ignorable modeling, which derives inference from the likelihood function based on a joint distribution of the variables and the missingness indicators, can be sensitive to misspecification of this distribution and may also have problems with identifying the parameters. Nonresponse two-phase sampling (NTS), which re-contacts and collects data from a subsample of the initial nonrespondents, has been used to reduce nonresponse bias. The additional data collected in phase II provides important information for identifying the parameters in the non-ignorable models. We propose a Bayesian selection model which utilizes the additional data from phase II and develop an efficient Markov chain Monte Carlo algorithm for the posterior computation. We illustrate the proposed model on simulation studies and a Quality of Life (QOL) dataset.

A Two-Phase Approach to Account for Unmeasured Confounding and Censoring of a Fixed Time Endpoint

Jaeun Choi, A. James O'Malley
Albert Einstein College of Medicine

Estimation of the effect of a treatment in the presence of unmeasured confounding is a common objective in observational studies. The Two Stage Least Squares (2SLS) Instrumental Variables (IV) procedure is frequently used but is not applicable to time-to-event data if some observations are censored. We develop a statistical method to account for unmeasured confounding of the effect of treatment on survival endpoints subject to censoring by considering censoring and confounding in sequence. We first jointly model survival time and treatment using a simultaneous equations model (SEM) under a specific bivariate distribution for the underlying data generating process. The joint model is used for the sole purpose of imputing the censored survival times. Then we apply an IV procedure to the completed dataset. This two-phase approach allows censoring to be accounted while preserving the robustness of the IV method to distributional miss-specifications in the joint model. The approach can be applicable to any type of survival outcome including continuous and fixed-time endpoints. The methodology is illustrated on two examples comparing endovascular vs open repair for patients with ruptured abdominal aortic aneurysm and reformulated vs original antidepressants for patients with major depression diagnosis. As the IV and the distributional assumptions cannot be jointly assessed from the observed data, we evaluate the sensitivity of the results to these assumptions.

Estimating Sexual Identity using the Annual Population Survey

Andrea Lacey, Bethan Russ
Office for National Statistics

Since 2015, the Office for National Statistics (ONS) has been asking UK adults about their self-perceived sexual identity on the Annual Population Survey (APS). Between 2012 and 2015 this information was collected using the ONS Integrated Household Survey, meaning four years of data are now available.

This poster will outline the design of the APS; including the sampling techniques and data collection methods implemented on the survey. We will focus in particular on how data on sexual identity is collected and the modes used to collect this data. The poster will also outline the estimation methods used to ensure our survey estimates are representative of the UK population, both overall and for specific domains. It will provide an overview of the sexual identity classification and provide some measures of precision on how robust estimates of sexual identity are, at the UK level and at local authority level.

The story of the million seeds – Using the "Rocket Science" project to engage the next generation in science and statistics

Ian Nevison, Emma Griffith, Libby Jackson
BioSS

As a profession we recognise the need not only to encourage new entrants but also to improve the statistical literacy of society at large and, in particular, current and future decision makers. Government also recognises the wider need to increase the numbers studying STEM subjects at school and university. It is hoped this will be furthered by the public interest in British ESA astronaut Tim Peake's recent mission to the International Space Station. The "Rocket Science" project seeks to capitalize on this by engaging both primary and secondary pupils in scientific experimentation, and statistics plays a key role.

Two kilograms of salad rocket seed spent six months at the International Space Station under microgravity prior to returning to earth. This was packaged up with other seed which had remained on earth. With over 5,000 schools receiving 100 seeds of each of the two types, the project involved the distribution and planting of over 1 million seeds. Participants each had to produce a randomised design and then grow both "space" and "earth" seed in order to compare outcomes such as the percentage of seeds that had germinated by the tenth day.

Teachers' resources were provided to schools to guide the design and analysis of their own experiment, along with materials explaining the relevant statistical aspects.

They subsequently entered their data into a national online database, enabling a statistical analysis to compare "space" and "earth" seed across all schools.

The statistical outreach potential is enormous. We are seeking to impart the importance of statistics in scientific experimentation, introduce statistical thinking and principles, cultivate critical evaluation of data and foster an interest in statistics.

The poster will demonstrate that a range of statistical concepts can be introduced through such a simple experiment and foster an understanding of good scientific practice.

Adaptive K-means Algorithm with New Distance Measure in Networks

Fatimah Almulhim, Peter Thwaites, Charles Taylor
University of Leeds

A network can be thought of as a group of discrete objects with relations between them, and representing these relations could be done using either a hierarchy or a lattice. So, networks have the ability to model various relations between data. Interestingly, graph theory is generally the best way to represent any network in multiple dimensions. Any network graph G = (V, E) consists of two elements, vertices V which represent the members in the network and edges E which represent the undirected links between vertices.

Any network graph has an underlying structure due to the heterogeneity of the nodes and edges. The idea of exploring this structure in terms of grouping the nodes is called clustering, the grouping is usually based on some similarity measure between nodes. In graph theory, applying clustering techniques is widely required in different applications e.g social network analysis. K-means is one of the popular method for cluster analysis in data mining, its goal is to partition V nodes into K clusters in which each node belongs to the cluster with the nearest centroid.

Using existing distance measure in graphs, the clustering process did not converge quickly due to lots of ties in the distances. Therefore, we are looking at a new way of measuring distances in graphs which is more suitable for clustering. Our idea is to measure the shortest distance between any two vertices by considering all possible paths length between them. A new distance metric is introduced which works efficiently with K-means clustering and produce more clustering solutions.

In addition, we adapt the K-means algorithm to be faster by around 5 times than the original algorithm by considering deterministic initial centroids at the first step.

Investigating the intraseasonal climate signal through Maximum Covariance Analysis: a case study for Tropical Brazil and India

Naurinete J. C. Barreto, Michel d. S. Mesquita, George U. Pedra, Jürgen Bader, Thomas Toniazzo, Saurabh Bhardwaj
Universidade Federal do Rio Grande do Norte, Brazil

Tropical rainfall time series present oscillations on a wide range of timescales, from hourly to decadal. In particular, such series are strongly linked with intraseasonal modulations at a scale between 20 to 100 days. Some of the common statistical methods that are used to capture this variability use Empirical Orthogonal Functions based on large-scale atmospheric variables, such as the outgoing longwave radiation or winds. Although this type of approach can reasonably capture the large-scale variability, it may not fully describe the complexity of local intraseasonal variability in tropical precipitation. We propose a new approach that is sensitive to both global and local processes affecting precipitation in a selected region. It is based on Maximum Covariance Analysis (MCA) applied to filtered daily anomalies of rainfall data against a group of covariates consisting of outgoing longwave radiation and the zonal component of the wind at the 850- and 200-hPa atmospheric levels. We provide numerical evidence that this approach captures the intraseasonal variability signal associated with rainfall, as well as the relationship between the seasonal and submonthly frequencies. The model proposed is applied to two regions: the Indian subcontinent and the tropical sector of Brazil. Over India, the first pattern of the MCA index (MCA1) and the second (MCA2) explain 45% and 26% of the intraseasonal variability, respectively. The most intense signal is captured from June to August, corresponding to the south-east monsoon. On submonthly timescales we find a maximum delay of 10 days between the values of MCA1 and MCA2. In tropical Brazil, the MCA1 and the MCA2 capture 68% and 15% of the variability, respectively. The signal is greatest between December and May, during the rainy season. The delay between the cycles of each MCA is up to 8 days.

An Investigation of the role of the Gamma Distribution in representing Positron Emission Tomography (PET) measurements

Tian Mou, Jian Huang, Eric Wolsztynski, Finbarr O'Sullivan
University College Cork

The Poisson-like structure of PET data with right tailed skewness and variability proportional to the mean is well appreciated in the field. The Gamma family of distributions can capture this structure. We use both real and simulated data to study applicability of the Gamma distribution to PET. The real data are from a series of dynamic phantom studies conducted by the American College of Radiology Imaging Network (ACRIN) as part of quality assurance for PET-based clinical trials [1]. Data from 43 different institutions are included in this series. Both traditional Fourier based and Maximum likelihood-based algorithms were used for reconstruction. The empirical analysis of the ACRIN data shows that for the likelihood-based reconstructed images, the errors in low count regions are better represented by the gamma distribution than by the normal distribution. The normal distribution is well justified for high count regions and traditional Fourier based reconstructions. These empirical findings are supported by simulation studies based 2-D Shepp-Vardi brain phantom. Both FBP and ML-EM reconstruction algorithms were used. Our findings are substantially in line with work reported by Teymurazyan et al. [2]. The Gamma structure has implications for inference including model fitting and diagnostics. The adaptation of standard analysis methods to accommodate the Gamma framework is straightforward and practical. While there are some gains associated using the gamma structure, the overall improvement in estimation efficiency is not dramatic.

References

[1] Scheuermann JS, et al: Qualification of PET Scanners for Use in Multicenter Cancer Clinical Trials: The American College of Radiology Imaging Network Experience. J Nucl Med*, 50(7), pp. 1187-1193, 2009.*

[2] A. Teymurazyan, et al: Properties of Noise in Positron Emission Tomography Images Reconstructed with Filtered-Backprojection and Row-Action Maximum Likelihood Algorithm. J Digit Imaging*, 26:447-456,2013.*

Spatial modelling of tumour features based on molecular imaging data

Eric Wolsztynski, Finbarr O'Sullivan, Janet O'Sullivan
University College Cork

Positron Emission Tomography (PET) is used routinely to quantify the uptake of a radiotracer, particularly 18F fluoro-deoxyglucose, within a cancer patient's body. PET offers a unique opportunity to image metabolic features of tumours, and complements MRI or CT information, which has a more physiological or anatomical focus. PET imaging modalities are used for non-invasive tumour characterisation, e.g. in terms of avidity or heterogeneity. Such variables can in turn be used in baseline prognostic models or for therapeutic effectiveness assessment, and statistically significant models enable patient-adaptive care.

We present recent developments in modelling the spatial distribution of tracer uptake within a volume of interest. The techniques considered are grounded in oncologic experience and offer the possibility to extend tumour characterisation beyond current routine summaries of PET imaging data. Typically, an idealised structure is fitted to the PET uptake observations where voxel data are expressed in terms of e.g. ellipsoidal or tubular coordinates with respect to the tumour core. Following this, the volumetric uptake distribution is described as a univariate function of parametrised radial location within the tumour. This voxel-level profiling may be fully- or semi-parametric and allows for localised assessment of tumour tissue texture and metabolic profile, including quantitation of the rate of metabolic change. We demonstrate how the implementation of a continuous, regularized spline-based evaluation of tubular PET uptake pattern yields valuable tumour characterisation, and how simpler constructs can also provide significant clinical summaries. Once variables of interest are derived, their relationship with patient outcome functionals is explored using longitudinal survival data analysis techniques.

We present various adaptations of such spatial modelling techniques, some parametric and some non-parametric, and illustrate their potential for treatment in a range of diseases including sarcoma, breast and lung cancers, based on US and Irish PET imaging datasets.

A Comparative Study on Filter and Wrapper Variable Selection Methods and Application to Metabolomics Data

Nurain Ibrahim, Marta (Garcia-Finana) Van Der Hoek, Gabriela Czanner, Lu-Yun Lian, Rudi Grosman
University of Liverpool

Metabolomics is a field of “omics” science that aims to study global metabolic changes in biological systems.It is largely based on the use of three techniques,namely Nuclear magnetic resonance (NMR) spectrometry,Gas chromatography mass spectrometry (GC-MS) and Liquid chromatography mass spectrometry (LC-MS).These techniques are applied to detect,identify and quantify small molecule metabolites from biological system & bio fluids with the ultimate goal of discriminating between groups of disease.The challenge is that metabolomics data is highly dimensional,p>>n where p is the number of variables and n is sample size and therefore variable selection is required in metabolomics studies.

The aim of our study is identify which variable selection methods perform better in metabolomics studies.In particular, we compare information gain and correlation based feature selection for filter methods together with sequential forward selection and sequential backward elimination for wrapper methods.We compare the variable selection methods on real metabolomics data that consists of 40 samples with 391 variables.The output variable is a categorical variable that represents three classes of diseases (wild type, mutant and complementary mutant).The data relates to a gram positive bacteria called *Staphylococcus aureus,*which is frequently found in the respiratory track and on the skin. Despite this bacteria is not always pathogenic,it is known to be a common cause of skin infections.

R programming software is used to analyse the data.Our result shows that, filter methods are performing better than wrapper methods based on the number of variables selected by these method.Filter methods just selected at most 7 best variables for this metabolomics data.The advantages of filter methods are that (i) they are faster than wrapper method,scalable,independent to the classifier and (ii) it is better computational complexity than wrapper methods.In conclusion,filter methods seem to be superior than wrapper methods when applying to metabolomics data.

Parameter estimation for BMRF models using a locally-linear simulation-based optimization approach

Wafa Almohri, Charles C Taylor, Robert G Aykroyd, Darren Treanor
University of Leeds and Taibah University

Maximum likelihood is a widely used method for parameter estimation in many models. However, when the likelihood is too expensive to evaluate, then estimation of parameters can be challenging. An exemplar problem is the Binary Markov Random Field (*BMRF*) model where the normalizing constant in the likelihood is time-consuming to compute as it involves a summation over a high dimensional sample space.

A new simulation-based method of estimating parameters for such models is introduced. This method uses summary statistics of the observed data and a “simulator box” whose output depends on a given set of model parameters. The method starts by choosing initial values for each parameter, then simulated data are produced from the "simulator box" in a sequential manner so that eventually they resemble the observed data. The parameter estimates are then given by the final set used in the simulation. The method assumes the relationship between the parameters and the summary statistics is locally linear. New design points are added and old ones removed in an adaptive manner, until convergence. This process solves a local linear model at each step, then as the number of simulations increases the best estimate of parameters is determined. In the poster the general framework will be presented which will be illustrated with an example dataset of real image data from a cancer classification study.

Open access to journal articles in dentistry: Prevalence and citation impact

Fang Hua, Helen Worthington, Tanya Walsh, Anne-Marie Glenny, Heyuan Sun
University of Manchester

Objectives: To investigate the current prevalence of open access (OA) in the field of dentistry, the means used to provide OA, as well as the association between OA and citation counts.

Methods: PubMed was searched for dental articles published in 2013. The OA status of each article was determined by manually checking Google, Google Scholar, PubMed and ResearchGate. Citation data were extracted from Google Scholar, Scopus and Web of Science. Chi-square tests were used to compare the OA prevalence by different subjects, study types, and continents of origin. The association between OA and citation count was studied with multivariable logistic regression analyses.

Results: A random sample of 908 articles was deemed eligible and therefore included. Among these, 416 were found freely available online, indicating an overall OA rate of 45.8%. Significant difference in OA rate was detected among articles in different subjects ($P < 0.001$) and among those from different continents ($P < 0.001$). Of articles that were OA, 74.2% were available via self-archiving ('Green road' OA), 53.3% were available from publishers ('Gold road' OA). According to multivariable logistic regression analyses, OA status was not significantly associated with either the existence of citation ($P = 0.37$) or the level of citation ($P = 0.52$).

Conclusions/clinical significance: In the field of dentistry, 54% of recent journal articles are behind the paywall (non-OA) one year after their publication dates. The 'Green road' of providing OA was more common than the 'Gold road'. No evidence suggested that OA articles received significantly more citations than non-OA articles.

A word of caution; using the Wold-Wolfowitz runs test to assess autocorrelation.

Lindesay Scott-Hayward, Monique Mackenzie
University of St Andrews

A Wold-Wolfowitz Runs test is commonly used to check for residual correlation. This work presents the efficacy of this test under various conditions and was assessed using simulation. Generalised Linear Models (GLM) were used with Poisson count data generated under a variety of signal-to-noise ratios and levels of correlation. The runs test compares the number of positive and negative sequences (runs) of residuals seen in the residuals to the number expected under independence. The test statistic is distributed N(0,1).

The runs test was found to return significant results more than expected by chance when the data are independent. Results varied from just above a nominal 5% error rate to the full 100% error rate depending on the data and signal-to-noise ratio (and specifications of the simulations). As a consequence of these findings, we have used an empirical approach to obtain more reliable p-values in place of the N(0,1) distribution, assumed to hold for the test statistic. Simulation results when using this empirical distribution returned a 5% chance of obtaining a significant result when the data were independent, in line with expectations. This result based on empirical distribution of the test statistic was maintained regardless of the signal-to-noise ratio. A reliable test for residual autocorrelation is important; the implications of assuming (incorrectly) that residuals are correlated means that inappropriate and more computationally expensive methods, such as Generalised Estimating Equations or Generalised Mixed Models, must be used (which typically increase computational time by 400%). This work is presented in the context of simulation-based power analysis for environmental impact assessment and where computational constraints are important.

The Creation and validation of a robust and brief psychometric tool to assess the challenge of living with cystic fibrosis: the CLCF-short form (CLCF-SF)

Gareth McCray, Holly Hope, Claire Glasscoe, Kevin Southern, Gillian Lancaster
Lancaster University

Objectives: Caring for a child with cystic fibrosis (CF) is a significant daily commitment for parent/caregivers. A tool that quantifies the challenge and the impact on the parent/caregivers would be useful, both clinically to monitor treatment burden and support families, and also as an important pragmatic outcome measure for clinical trials. Focus groups, cognitive interviews and the collation and analysis of expert opinion were used to create the Caring for Children with Cystic Fibrosis (CLCF) questionnaire, which consists of 249 items. However, logistically this questionnaire may be difficult to utilise given its length. The aim of this study was to create and validate a condensed set of items, the CLCF-Short Form, which best reflect the construct intended to be measured by the full questionnaire.

Methods: A genetic algorithm is an iterative technique used for the optimisation of problems which operates via evolution of increasingly apt solutions to those problems. This kind of algorithm was used in this study to select a subset of items which best measured the central construct.

Results: The measure as a whole had good internal reliability with a Cronbach's alpha of 0.82 (95% CI 0.77, 0.88). Convergent validity of the measure was demonstrated through correlations with BDI (0.48), STAI State measure (0.41), STAI Trait measure (0.43), summed treatment score (0.18), and model-based scores of caregiver treatment management (0.49) and child treatment management (0.46). The CLCF short form (CLCF-SF) provides a good psychometric tool with a reduced number of items that reflects the key information gained from the original long form well. The CLCF-SF has excellent internal consistency and contextual validity. The tool has great promise as a robust measure for use both in a clinical context and as a pragmatic outcome measure for interventional studies that involve caring for children with CF.

The National Audit Office: Using data to help the nation spend wisely

Heather Reeve-Black, Floria Hau, Robindra Neogi, Leanne Stickland, Mark Edward, Rose Martin
National Audit Office

The National Audit Office (NAO) scrutinises public spending on behalf of Parliament.

Each year we publish around 60 value for money studies, examining whether government departments have used public money efficiently, effectively and with economy. We aim to hold government to account for how it uses public money, and to drive lasting improvement in public services.

Our value for money studies are evidence based and we draw our conclusions on the basis of rigorous analysis.

The NAO's statistics community of practice is dedicated to promoting high standards and expanding our knowledge, both in terms of developing our analytical capability and making statistics meaningful for Parliament and the public. We would like the opportunity to showcase some of our work, including:

The Future of Jobcentre Plus

To support the inquiry into the future of Jobcentre Plus, we briefed the Work and Pensions Committee on recent trends in unemployment benefit caseload, and regional variation in claimant characteristics and accessibility by public transport.

Financial sustainability of local authorities: capital expenditure and resourcing

We examined recent changes in patterns of capital spending and financing in local authorities, focussing on the implications and risks to financial and service sustainability. To extend the impact of the report we created an interactive visualization that allows users to explore the main themes.

Maternity services in England

We used a list of maternity service locations to calculate average drive times to the nearest of each type of maternity service for every LSOA in England. We recommended that "*clinical commissioning groups and trusts should agree long-term, sustainable plans for the distribution and capacity of maternity services in their locality*".

Estimating income using consumption expenditure variables for urban households of Costa Rica

Alejandra Arias
University of Costa Rica

Household surveys are currently the main method for collecting data on household income in Costa Rica and many developing countries, however, an optimal income data capture is difficult, either because insecurity or lack of knowledge of the informants and this could cause inaccurate and erroneous results. The main objective of this article is to present a simple and faster alternative to estimate household income based on consumption expenditure data for urban households, without requesting sensitive data to informants. This study is carried out using data from the National Survey of Income and Expenses of Costa Rica, 2013. The analysis compares three multiple regression models to assess statistical relevance and applicability in future household surveys. The first model is an updated indicator used for stratification of households and segments of census 2000, it is constructed based on information of the housing and the head of the household; the other two models are constructed based mostly on household consumption expenditure of last month: a dichotomous model (expenses in private health services, household services, cable or satellite television and insurance expenses) and a continuous model (amount of expenditure in private health services, insurance, electricity and telephone).

In terms of good of fit, all three models are similar. However, in assessing the complexity of the indicators construction, dichotomous model is the simplest one and for this reason, is considered more appropriate. Because the first model had been raised for stratification of households and segments, another important result is how this indicator should be updated in order of implement it in coming censuses. Validation of these three models was performed by comparing income deciles between original and estimated income, obtaining very similar results.

Modelling with Algebraic Techniques

Shaoxiong Hu, Hugo Maruri-Aguilar
Queen Mary, University of London

An important class of regression models both from a scientific and an algebraic point of view are hierarchical models. In a hierarchical model, a term such as x_1x_2 can only appear if the terms x_1, x_2 and the intercept are also included in the model . Such models are described as staircase models in algebra because of this divisibility property. Hierarchical models may be identified via the algebraic method in experimental design.

A description of terms in a hierarchical model degree by degree is performed by the Hilbert function. In such a study, {\em Betti numbers} effectively count those monomials. A special type of models, called lex segment model which has maximal Betti numbers. We work around an algorithm that yields corner cut models that are also lex segment models. We can get a sufficient condition for this construction.

We propose a modelling strategy that applies the conditions we have found to select models in a dataset from the literature. The algorithm above can be adapted to explore different regression models. We focus on corner cut models described above and compare them using traditional measures such as R^2, the adjusted one, R_A^2 and the residual error which exactly indicates how close the data are to the fitted regression. We will embed our algebraic methodology with model selection techniques such as lasso. We compare the algebraic methodology to statistical model selection methods using both simulated data and cancer data from literature (Tibshirani 1996). Future work will find out whether the model selection can be meaningfully informed by topological features.

Kinetic Analysis of PET Imaging Radiotracers with Metabolite Adjustment

Sarah Murphy, Eric Wolsztynski, Mark Muzi, Finbarr O'Sullivan
University College Cork

Positron emission tomography (PET) is widely used in the clinical management of cancer patients. While the most common radiotracer is 18F-fluorodeoxyglucose, other tracers are also being evaluated. If PET imaging studies are acquired in dynamic mode, kinetic analysis is utilised in order to recover sophisticated metabolic information about the local tissue from the data. Most often compartmental models are used for this purpose but nonparametric generalizations of these are also possible (Hawe et al.). When the injected PET radiotracer leads to production of recirculating metabolites, an understanding of the kinetics of these metabolites may be key to interpreting these data.

We present examples and explore modeling approaches for this situation. One example arises in the used of 11C-thymidine (TdR) as a radiotracer of cell proliferation and DNA synthesis. Data from brain and somatic tumors are considered. A nonparametric, multiple-input model of the tracer kinetics, for the quantification of dynamic TdR data is proposed. This model relies on a piecewise constant function of the time course information. Its calibration yields a data-adaptive representation of the retention of tracer in tissue (Hawe et al.), as opposed to compartmental models which impose a constrained shape (Wells et al.). Simulations were undertaken to examine the efficacy of the multiple-input compartmental and nonparametric models and also to understand of the convergence rates of kinetic parameters of interest. Critical comparisons between the models are performed using cross validation.

References

Hawe, D. et al. *(2012). Kinetic analysis of dynamic positron emission tomography data using open source image processing and statistical inference tools. WIRES Comput Stat,* ***4(3)****: 315-322.*

Wells, J.M. et al. *(2002). Kinetic analysis of 2-[11C] thymidine PET imaging studies of malignant brain tumors: Compartmental Model Investigation and Mathematical Analysis. Mol Imaging.* ***1(3)****: 151–159.*

Supported by SFI-PI11/1027 and NCIP01-CA42045

Models for forensic speaker comparison

Tereza Neocleous
University of Glasgow

We present preliminary results from a collaboration between statisticians and forensic phoneticians which aims to develop models that quantify forensic phonetic evidence more reliably. Focusing initially on the single hesitation marker "um" from speech recordings from the DyViS database (Nolan et al 2009), we analyse linguistic features (vowel formants F1, F2, F3 and duration) in univariate and multivariate random effect models. Simulation studies show that, with a sufficient number of "um" tokens, speaker comparisons can be made with relatively low error rates. While this result is promising for single vowel analysis, in a realistic scenario phoneticians would need to analyse all words/vowels from a recording simultaneously. We discuss the statistical challenges involved in combining evidence from multiple words/vowels and some possibilities for dealing with this issue.

Methods of Predicting Football Match Results and Final League Positions Using "Phases of Play", In-Play Odds and Monte Carlo Simulations

Gordon Hunter, Ishan Rashid, Adam Bartholomeusz
Kingston University

Association Football (or "Soccer") attracts huge interest and, in some cases, fanatical following from supporters all around the world. Over recent years, gambling on match outcomes and overall end of season league positions have become very popular, and assisted in the major growth of the gambling industry, creating a substantial number of new jobs.

In this paper, we discuss how statistical and probabilistic models can be used to predict such things. We focus on two main tasks - predicting outcomes of individual matches during the same game based on features for different "Phases of Play" (Bedford and Baglin, 2009) and relating these to in-play match odds, and predicting both match and end of season league postions based on Monte Carlo simulations, bearing in mind the attacking and defensive records of both sides. In both cases, real match and/or league results data are used to train the models used, and the predicted outcomes made by the models compared with what actually happened.

Comparing Effects of Biologic Agents in Treating Patients with Rheumatoid Arthritis

Ingunn Fride Tvete, Bent Natvig, Jørund Gåsemyr, Nils Meland, Marianne Røine, Marianne Klemp
The Norwegian Computing Center

Objective

Rheumatoid arthritis patients are treated with disease modifying anti-rheumatic drugs (DMARDs) and the newer, more expensive, biologic drugs. We sought to compare and rank the biologics with respect to efficacy.

Methods

A literature search identified 54 publications considering altogether 9 biologics (abatacept, adalimumab, anakinra, certolizumab, etanercept, golimumab, infliximab, rituximab and tocilizumab). We developed a multiple treatment comparison regression model allowing the number experiencing a 50% improvement on the American College of Rheumatology scale to be dependent upon dose level and disease duration for comparing the relative effect of biologics. Unlike others, we distinguished between joint DMARD and biologic or exclusively biologic drug use, and allowed this difference to be drug dependent.

Results

The drug effect was dependent on dose level, but not on disease duration, with higher doses indicating better effects for all drugs. All biologic agents were effective compared to placebo, with certolizumab the most effective and adalimumab (without DMARD treatment) and adalimumab/ etanercept (combined with DMARD treatment) the least effective. The drugs were in general more effective, except for etanercept, when given together with DMARDs.

Conclusions

All biologic agents were effective compared to placebo, with certolizumab the most effective. Several published comparison analyses have presented somewhat diverging results. The final result depends on the chosen model, trials included, handling of heterogeneity, possibly background information considered and if and how joint treatment with other drugs are considered. We believe our approach, including all relevant comparisons, both direct and indirect, while specifying whether the drugs were given alone or in combination with DMARDs, is an important contribution in the biologic drugs comparison analyses literature.

Impact of Real-World data quality on ability to detect safety signals in post market surveillance.

Rene Beattie, Thomas Marshall, Joost de Folter, Mark Trusheim, Aiden Flynn
Exploristics Ltd

Clinical trials undertaken during development of a new drug or treatment have limited scope to detect relatively uncommon side effects that would be observable in the general population. Once a drug is distributed to a wider population, this is a less controlled environment and the quality of safety data that can be collated is uncertain. The objective of this study is to estimate the robustness of safety signal detection to some data quality factors.

The Kerus clinical trial simulator was used to simulate the alternative approaches for safety signal detection:

- Extended clinical trials
- A registry trial for patients post-authorisation.
- A pragmatic trial for patients post-authorisation compared to a population receiving the current best standard of care.

The study was based on published clinical trials (to obtain example drug efficacy for an authorised drug and associated cardiac event data), supplemented with pertinent therapeutic and study conduct information obtained from public sources. The impact of data quality was studied by varying the level of patient recruitment and the reporting of existing interacting health conditions.

Statistical analysis of simulated patient populations indicated extension of the clinical trials does not appear to be an economically viable approach to improving safety information. However, in a perfectly reported registry trial (without any real-world effects modelled) the safety signal would have been reported as early as 3 months post-authorisation. An incomplete registry (50% of treated patients included, of which only 50% report their health history) would be expected to detect adverse cardiac events within 12 months. A comparison with pragmatic studies show that a registry would perform similarly under the assumptions used. By adjusting parameters affecting data quality (bias, missing data etc.) it is possible to estimate the level of rigour required in collating real world data in order to achieve statistically significant outcomes in real-world studies.

Preliminary Study of Medical Cost Prediction: Target Disease Expansion in the UK for Using a Bayesian network

Shuntaro Yui, Shinji Tarumi, Hajime Sasaki, Takanobu Osaki, Hideyuki Ban, Norman Stein, Sheila McCorkindale, Martin Gibson
Hitachi Europe

Currently 10% of the NHS budget is spent on diabetes treatment and this is likely to rise to at least 17% by 2035, which is clearly unsustainable. Prevention could be effective for reducing cost burden though it takes much time to get the quantitative impact. In order to demonstrate the degree to which delaying or preventing the onset of type 2 diabetes could be economically effective, we have been establishing a cost modelling framework.

Salford has a population of 242,000 and an integrated electronic 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. We had developed the above framework based on a Markov model; however, this framework faces two challenges 1) inability to predict for those whose indicators are missing, and 2) high leverage for target disease expansion.

In order to solve the challenges above, we propose a new cost prediction method using a Bayesian network with automatic discretised function which converts from continuous values to discrete ones. The new method could interpolate significant missing data values because the new model describes the complex relationship between each indication such as test results and disease status. It also introduces a data-driven approach to model construction because it can achieve automatic model construction by learning from data.

Experimental results using Salford diabetes data (14169 patients) showed that the amount of data available was three times larger because of missing data inclusion and less leverage of disease expansion, whilst predicted medical cost was £850 in which prediction error achieved less than 5%. It demonstrates that the proposed method can achieve target disease expansion without complicated efforts by interpolation of missing data values and automatic model construction almost without clinical experts' knowledge.

Bayesian approach for clustered interval-censored data with time-varying covariate effects

Bin Zhang, Yue Zhang, Xia Wang
Cincinnati Children's Hospital Medical Center

Interval-censored data arise when failure times cannot be observed exactly but can only be determined to lie within an interval. interval-censored data are very common in clinical trials and epidemiological studies. In this study, we consider a Bayesian approach for correlated interval-censored data under a dynamic Cox regression model. Some methods that incorporate right censoring have been developed for clustered data with temporal covariate effects. However, no interval-censored data analysis has been considered under the same circumstance to the best of authors' knowledge. In this paper, we estimate piecewise constant coefficients based on a dynamic Cox regression model under Bayesian framework. The dimensions of coefficients are automatically determined by reversible jump Markov chain Monte Carlo algorithm. Meanwhile, we use a shared frailty factor for unobserved heterogeneity or for statistical dependence between observations. Simulation studies are conducted to evaluate the performance of proposed method. The methodology is exemplified with a pediatric study on children's dental health data.

Calibrating a Whole Body Circulation Model for Image-Based AIF Extraction in the context of Blood Flow-Metabolism Mismatch Assessment in LABCa Patients

Zhaoyan Xiu, Jian Huang, Janet O'Sullivan, Eric Wolsztynski, Mankoff Dave, Finbarr O'Sullivan
University College Cork

Blood flow-metabolism mismatch from PET studies with 15O-H2O (water) and 18F-FDG (FDG) is a promising diagnostic for monitoring treatment in locally advanced breast cancer (LABCa) patients. The mismatch measurement requires kinetic analysis with the arterial blood signal (AIF) as an input function. Arterial sampling is invasive and not always feasible. We have developed a whole body circulation model for representation of the circulation of a tracer atom in the body [1]. Based on fitting this model to time courses from multiple Region of Interests (ROIs) on same dynamic study simultaneously, this work describes a novel statistical method for image based AIF extraction. A penalized nonlinear least squares optimization is implemented for estimation of the model parameters. 53 LABCa patients who had dynamic PET studies with water and FDG were used to explore the performance of the proposed method. For each PET study two AIFs were recovered, one using the automated statistical method and a manually extracted one derived from a region of interest placed over the left-ventricle (LV-ROI). Flow-metabolism mismatch were obtained with each AIF and kinetic and prognostic reliability comparisons made. Strong correlations were found between kinetic assessments produced by both AIFs. The statistically extracted AIFs retained the full prognostic value, for pathologic response and overall survival, of LV-ROI AIFs. Our work explores the ability of the circulation model to simultaneously describe PET-measured time-course signals in key structures such as the liver, lung and right-ventricle of the heart. The results provide a basis for adapting the circulatory model to PET radiotracers in which there may be more limited opportunity to use directly sampled arterial measurements for model calibration.

[1] J. Huang and F. O'Sullivan *An analysis of whole body tracer kinetics in dynamic PET studies with application to image-based blood input function extraction*. IEEE Trans Med Imaging 2014 May;33(5):1093-108.

Incorporating Single Treatment Arm Evidence into a Network Meta Analysis (if you must!)

Joy Leahy, Cathal Walsh
Trinity College Dublin

Combining all available evidence in a Network Meta Analysis (NMA) is important for facilitating the decision making process surrounding the appropriate choice of treatment regimens in a clinical setting. Randomised Control Trials (RCTs) are considered the gold standard of evidence, as potential bias is minimised. However, much of the evidence available can be from other sources, such as one-armed, single-agent and observational studies. These can contain valuable information, so it is necessary to examine approaches for including them in NMAs.

We propose including single-agent trials by choosing a similar arm from another trial in the network to use as a comparator (matched) arm. By simulating trials where the effects of treatments are known, and sampling across the potential matches, we vary parameters which are likely to influence the effectiveness of matching. Parameters examined are the standard deviation of the between study effect, the effect of covariates, the treatment effect, the number of patients, and the size of the between study type effect. The objective is to assess which parameters influence the effectiveness of matching, analyse methods for choosing matched arms, and assess whether they are likely to work better than using RCT evidence alone. We apply these techniques to a hepatitis C dataset in treatment naive patients.

Matching by the covariate generally produces better estimates than randomly matching. When each trial study effect is set to zero, including single-agent evidence produces better estimates than including only RCT. Increasing the standard deviation of the between study effect gradually worsens the estimates to a point where including single-agent evidence produces less accurate estimates.

Under certain conditions, including single-agent evidence can increase the accuracy of our estimates of treatment effects in an NMA. However, we must exercise caution when including single-agent estimates as they may introduce bias into the model.

Spatial analysis of ecosystem risks of complex soil pollution in flood affected areas in the Czech Republic

Jan Skála
Research Institute for Soil and Water Conservation, Czech University of Life Sciences Prague

A rigorous analysis of soil contamination in floodplain soils in the Czech Republic was conducted by an extensive soil sampling (100 soil profiles) and screening assessment of ecosystem risks combined with a multidimensional statistical analysis and spatial analysis of ecosystem risks of soil contamination with the objectives of revealing both the spatial patterns of pollution magnitude as well as pollutants' composition.

There were estimated ecosystem risks of soil pollution by trace elements (As, Cd, Cu, Hg, Pb, Ni, Zn) and persistent organic pollutants (PCBs, PAHs, organochlorine pesticides – DDTs, HCHs) based on a confrontation of environmental concentrations with referenced values for potential ecosystem risks of soil contamination. Relative contributions of each element/substance to the estimation of ecosystem risks (overall hazard index) were calculated and proportional similarities between all the localities were calculated. We correlated the proportional similarities matrix with various matrixes by the Mantel correlation test. We tested similarity matrix based on soil characteristics and various spatial configuration matrixes. The results implied that spatial connections and potential sources affinity were main drivers of similarity and the observed pollution profiles are rather due to the input of contaminants from sources than to the specific accumulation and storage properties of the studied soils. A multivariate Mantel correlogram was than computed to analyse the spatial correlation of proportional similarities among the samples.

Relative Efficiency and Sensitivity of Latin Square Design

Abimibola Oladugba, Leonard Nwogu-Ikojo
University of Nigeria, Nsukka

This work examines the relative efficiency and sensitivity of Latin Square Design (LSD) to Randomized Complete Block Design (RCBD). The result shows that LSD is better than RCBD in terms of relative efficiency, which depends on the error variances of the two designs. While in under the relative sensitivity the RCBD tends to give more information than the LSD in some cases; since the relative sensitivity depends on the error variances and the number of degrees of freedom for error.

Interaction between state's political chracteristics and regime types with natural hazards

Md Rezwan Siddiqui, Dr Helen Adams
King's College London

Natural hazard does not occur in isolation rather in conjuncture of social, economic and political space. Hazard loss and fatalities are affected by the elements of state polity and regimes characteristics. The objective is to explore how different regime types interact with natural hazards. This discussion is not aimed to present any collective theory of relationship between disaster and politics, rather to shed light on unexplored variables through indicator based analysis. Data from POLTY IV, EM-DAT, AidData and WDI has been used for 161 countries from 1980-2015. Economic loss, disaster mortality and number of people affected are the outcome variables. Neural network analysis, automatic linear modelling and ANOVA are used to identify relative significance and level of influence of state's political characteristics on natural disaster outcome.

Disaster outcomes significantly vary based on the types and characteristics of regimes. Polity score, regime's durability shows significant positive correlation relationship with disaster outcomes. Financial damage and the number of affected people changes significantly with the alternation of regimes types from full autocracy to full democracy. State's stability, bring decreasing impact on disaster fatality and the number of people affected. This explains the importance of regime's stability in disaster risk reduction. Instable regimes are more vulnerable to high disaster losses. Regime types are also highly correlated with state income level. High level of income is found in countries with full democracy and vice versa, which can cause more financial damage than other types of regime.

Exploration and explanation of these relationships could have significance positive impact in DRR through intervention to strengthen state's political structure. Along with economic development, guided political development could be useful tool for DRR.

Photocatalytic destruction of azo dyes: co-relation between experimental and theoretical approaches by Markov Chain Monte Carlo simulation

Vivek Verma, Priyadarshi Roy Chowdhury MRSC, Krishna G. Bhattacharyya
Department of Statistics, Gauhati University, India

3:1 Fe/Ti layered double hydroxide (LDH) nanoparticles, synthesized by a single step hydrothermal method, exhibited excellent semiconductor properties with remarkable visible light decolorization potential for both Crystal Violet and Reactive Orange 16 respectively, the azo dyes being mostly present in industrial waste water, The photodegradations of the dyes, by the LDH, is found to be higher than that observed with commercial catalysts like FeO, Fe_2O_3, TiO_2 and Degussa P25. The photocatalysis proceeds through e^--h^+ hopping and by dye photosensitized mechanistic pathways. The dynamics of electron transport through the semiconductor Fe/Ti LDH nanomaterial is investigated through Markov Chain Monte Carlo (MCMC) simulations. As a result of co-relation between simulated and the characterization results, associated with electron transport, across the LDH, it could be validated that electron transport is the principal factor which imparts high photoactivity to semiconducting nanomaterial. This work provides a new insight on photocatalysis and charge carrier dynamics of Fe/Ti LDH nanoparticles under random set up.

REFERENCES

1. P. Roy Chowdhury and K.G. Bhattacharyya, Dalton Trans., *2015,* ***44,*** *6809.*

2. P. Roy Chowdhury and K.G. Bhattacharyya, RSC Adv., *2015,* ***5,*** *92189.*

3. S. Chib and E. Greenberg, Am. Stat. J., *1995,* ***49,*** *327.*

INDEX

H

I

J

K

L

M

N

O

P

R

S

T

V

W

X

Y

Z