2021

RSC

LANCASTER

44th Research Students' Conference in Probability and Statistics

Conference Proceedings
27th-29th July 2021
Lancaster University

Welcome Address

Hello and welcome to the 44th Research Students' Conference in Probability and Statistics a.k.a. RSC2021!

The aim of the conference is to highlight and celebrate the quality of research carried out by PhD students in statistics and probability. We want for students to be able to share their work and experience in a friendly environment whilst maintaining a solid standard of content. The upside of the hosting a virtual nature of the conference is that it is accessible to a far more spread out demographic. This year we are joined by students coming from 65 universities, in 22 countries.

We hope you find the programme interesting; the talks cover a wide range of topics so there is something for everyone. Additionally, there will be three keynote talks by world-renown researchers which will showcases recent advances in the respective fields.

On Wednesday, we will have a session of π-minute theses. These will be a series of lightning talks where speakers will try to convey the key aspects of their PhD projects in 3 minutes and 14 seconds (or 3.14 minutes, this is a very contentious issue).

We would like take this opportunity to thank the keynote speakers, Prof. Peter Diggle, Prof. Kerrie Mengersen and Prof. Mark Girolami for taking the time to join us. Additionally, we wish to extend our thanks to the Royal Statistical Society, the RSS Lancashire and Cumbria local group, and the STOR-i Centre for Doctoral Training, in particular Prof. Jonathan Tawn and Wendy Shimmin, for the support in organising the conference.

We sincerely hope that you enjoy your time during the three days!

Best wishes,
The RSC 2021 Organising Committee

Contents

Timetable of Events

Time (BST)	Tuesday (27th July)	Wednesday (28th July)	Thursday (29th July)
09:00		**Keynote: Kerrie Mengersen** *From meat pies to spaghetti bolognese*	
09:15			
09:30			
09:45			Parallel sessions: Probability Statistical Modelling I
10:00		*Coffee*	
10:15		Parallel sessions: Computational Statistics Epidemiology	
10:30			
10:45			*Coffee*
11:00			**Keynote: Mark Girolami** *The Statistical Finite Element Method*
11:15		*Coffee*	
11:30		Single session: Model Selection II	
11:45			
12:00			*Coffee*
12:15			Single session: Statistical Modelling II
12:30	**Welcome**	*Lunch*	
12:45			
13:00	**Keynote: Peter Diggle** *Geostatistics, Point process and COVID-19 monitoring*	Single session: Medical Statistics	
13:15			
13:30			
13:45			
14:00	*Coffee*		
14:15	Parallel sessions: Clinical Trials Model Selection I		
14:30			
14:45			
15:00			
15:15	*Coffee*		
15:30	Single session: Bayesian Statistics		
15:45			
16:00			
16:15			
16:30		**π-minute Theses & Quiz**	
16:45			
17:00			
17:15			
17:30			
17:45			
18:00			

Helpful Information

Please note that, all the times given throughout this document are in the British Summer Time (BST). If you are outside of the UK, make sure that you correctly convert the times.

Contact Details

Please direct your inquiries to rsc2021@lancaster.ac.uk. You can also message any of the admins on the conference's Slack.

Attending talks

By now everyone should have received an email with a link to the RSC2021 Slack workspace. This is the conference central where we will post links to all sessions. If you haven't got access to the Slack, please contact the committee. All of the talks will take place via Microsoft Teams and so for this reason we recommend everyone the desktop app if possible (https://www.microsoft.com/en-ie/microsoft-teams/download-app). When attending the talks, if you wish ask the speaker a question you will need to post it in the live chat and wait for the session chair to pick it out.

Giving talks

If you are presenting a 15-minute talk you will receive an email with a presenter link. During the talk you will be sharing your own screen via Microsoft Teams. If you are having bandwidth issues, turning off your camera stream can help. Additional 5 minutes are allocated for questions. These will be posed in the live chat and read out by the chair. If the questions are taking too long, we advise you move the discussion to Slack.

If you are presenting during the π-minute thesis session, you will need to send one slide, in the PDF format, to the RSC Committee. The slide is optional; you can present without any aids if you wish. During the session, the chair will be sharing the slides.

Chairing talks

If you are chairing a session, you will receive the relevant link via email. We advise you familiarise yourself with the talks in your session. To ensure the conference runs smoothly, we advise all chairs to go into the session 5 minutes prior to the scheduled start. If one of the speakers is having technical difficulties, we suggest you move to the next scheduled talk and come back to the speaker at the end of the session when the issues have hopefully been resolved. We aim to have at least one RSC Committee member in each session to offer help.

If for some reason you are no longer able to chair, please contact the committee as soon as possible.

Keynote Speakers

Geostatistics, Point Process and COVID-19 Monitoring

Date: 27th July
Day: Tuesday
Time: 13:00-14:00
Room: Keynote I

Professor Peter Diggle
CHICAS, Lancaster University Medical School

In studies of the geographical variation in disease prevalence, data are often presented as a set of case-counts at each of a discrete set of fixed locations or spatial sampling units, and modelled accordingly. In this talk, I will argue that it can be advantageous to derive a model for spatially discrete count data from an underlying spatially continuous point process of individual cases. I will discuss some practical implication and describe how we have used point process models to monitor aspects of the COVID-19 epidemic in the UK.

Diggle, P.J., Moraga, P., Rowlingson, B. and Taylor, B. (2013). Spatial and spatiotemporal log-Gaussian Cox processes: extending the geostatistical paradigm. *Statistical Science*, **28**, 542-563.

Fry, R.J., Hollinghurst, J., Stagg, H.R., Thompson, D.A., Fronterre, C., Orton, C., Lyons, R.A., Ford, D.V., Sheikh, A. and Diggle, P.J. (2021). Real-time spatial health surveillance: mapping the UK COVID-19 epidemic. *International Journal of Medical Informatics*, `DOI:10.1016/j.ijmedinf.2021.104400`

Peter Diggle is Distinguished University Professor of Statistics in CHICAS, a teaching and research group within the Lancaster Medical School at Lancaster University working at the interface of statistics, epidemiology and health informatics. He is also the Director of Training for Health Data Research UK. He holds a post at the University of Liverpool Department of Epidemiology and Population Health, and adjunct appointments at Johns Hopkins University, Columbia University, and Yale University. Professor Diggle's research involves the development of statistical methods for spatial and longitudinal data analysis, motivated by applications in the biomedical, health and environmental sciences. He has published 12 books and more than 250 articles on these topics in the open literature.

He was awarded the Royal Statistical Society's Guy Medal in Silver in 1997, and is a former editor of the Society's Journal, Series B. Professor Diggle was also president of the Royal Statistical Society (2014-2016). He was founding co-editor of the journal Biostatistics between 1999 and 2009, and is a trustee for the Biometrika Trust. He has served the UK Medical Research Council as a member of their Population and Systems Medicine Research Board and as chair of their Skills Development Panel.

Date: 28th July
Day: Wednesday
Time: 09:00-10:00
Room: Keynote II

From Meat Pies to Spaghetti Bolognese

Professor Kerrie Mengersen
Queensland University of Technology

Sometimes we undertake statistical analysis with a clear purpose in mind, access to all the necessary data, an obvious statistical model and ample computational resources. Often we don't. In this presentation, I would like to explore what happens when we substitute ingredients in cooking up our analyses. In particular, I will touch on some of the topics in Bayesian statistics that are of current interest, including the inclusion and impact of prior information, quantifying and communicating statistical uncertainty, and personalised inference. I would also like to touch on how the same analogy can apply to our career trajectories.

Kerrie Mengersen is Distinguished Professor in the School of Mathematical Sciences at Queensland University of Technology. Professor Mengersen's research focuses on using and developing new statistical and computational methods that can help to solve complex problems in the real world. These problems are in the fields of environment, genetics, health and medicine, and industry. Her research interests include complex systems modelling, Bayesian statistical modelling, Bayesian networks, and applied statistics. She has published over 350 refereed journal publications on these topics.

Professor Mengersen is acknowledged to be one of the leading researchers in her discipline. She also received two prestigious awards in 2016: the Statistical Society of Australia's Pitman Medal, the highest award presented by the Society and the first time it has been presented to a woman, and the Research Excellence award by the Cooperative Research Centre for Spatial Analysis (CRCSI). In 2018 Professor Mengersen was elected a Fellow of the Australian Academy of Science (AAS); a Fellow of the Academy of Social Sciences in Australia (ASSA); and an Invited Fellow of the Queensland Academy of Arts and Sciences (QAAS).

The Statistical Finite Element Method

Date: 29th July
Day: Thursday
Time: 11:00-12:00
Room: Keynote III

Professor Mark Girolami
University of Cambridge

Science and engineering have benefited greatly from the ability of finite element methods (FEMs) to simulate nonlinear, time-dependent complex systems. The recent advent of extensive data collection from such systems now raises the question of how to systematically incorporate these data into finite element models, consistently updating the solution in the face of mathematical model misspecification with physical reality. This article describes general and widely applicable statistical methodology for the coherent synthesis of data with FEM models, providing a data-driven probability distribution that captures all sources of uncertainty in the pairing of FEM with measurements.

Mark Girolami holds the Sir Kirby Laing Professorship of Civil Engineering within the Department of Engineering at the University of Cambridge, where he also holds the Royal Academy of Engineering Research Chair in Data Centric Engineering. Professor Girolami was also one of the original founding Executive Directors of the Alan Turing Institute the UK's national institute for Data Science and Artificial Intelligence, after which he was appointed as Strategic Programme Director at Turing, where he established and continues to lead the Lloyd's Register Foundation Programme on Data Centric Engineering. His research interests include computational statistics, Bayesian statistical methodology, and applications of probabilistic, stochastic and statistical modelling in the engineering and natural sciences.

Professor Girolami is a fellow of the Royal Society of Edinburgh, an EPSRC Advanced Research Fellow (2007-2012), an EPSRC Established Career Research Fellow (2012-2018), and a recipient of a Royal Society Wolfson Research Merit Award. He delivered the IMS Medallion Lecture at the Joint Statistical Meeting 2017, and the Bernoulli Society Forum Lecture at the European Meeting of Statisticians 2017. Professor Girolami currently serves as the Editor-in-Chief of Statistics and Computing, and Data Centric Engineering.

Schedule of Talks

Keynote I	**Tuesday 27th July, 13:00-14:00**
Chair: Jordan Richards	
Speaker	**Title**
Professor Peter Diggle	Geostatistics, Point Process and Covid-19 Monitoring

Clinical Trials	**Tuesday 27th July, 14:15-15:15**
Chair: Callum Barltrop	
Speaker	**Title**
Hanan Alzahrani	Optimal Design of Clinical Trials in Multi-state Models
A. Haris Jameel	Assessing Patient Benefit with Semi-Markov Multi-State Models
Jinran Zhan	Comparison of Frequentist and Bayesian Methods for Borrowing Historical Two-Arm Data in Clinical Trials

Model Selection I	**Tuesday 27th July, 14:15-15:15**
Chair: Jack Kennedy	
Speaker	**Title**
Andrew McInerney	Information-criteria-based Model Selection for Neural Networks
Laura McQuaid	Penalized Multi-Parameter Regression Modelling
Meadhbh O'Neill	Smooth BIC Variable Selection Procedure for Heteroscedastic Data

Bayesian Statistics	**Tuesday 27th July, 15:30-16:30**
Chair: Szymon Urbas	
Speaker	**Title**
Hannah Comiskey	Estimating the Proportion of Modern Contraceptive Methods supplied by the Public and Private Sectors using a Bayesian Hierarchical Penalized Spline Model
Mai Dao	Bayesian Variable Selection in Binary Quantile Regression using a Quantile-specific Prior
Jack Kennedy	Decision Support for Large Offshore Wind Farms via Bayesian Optimisation

Keynote II	**Wednesday 28th July, 9:00-10:00**
Chair: Srshti Putcha	
Speaker	**Title**
Professor Kerrie Mengersen	From Meat Pies to Spaghetti Bolognese

Computational Statistics	**Wednesday 28th July, 10:15-11:15**
Chair: Tom Grundy	
Speaker	**Title**
Muhammad Mahmudul Hasan	Bayes Linear Emulation for a Complicated Computer Simulation
Virgilio Pérez	How to Increase the Sample Size: The use of Rolling Windows
Hayley Smith	A Review of Simulation Studies Comparing Statistical and Machine Learning Approaches to Risk Prediction for Time-to-Event Data

Epidemiology	**Wednesday 28th July, 10:15-11:15**
Chair: Samantha Ofili	
Speaker	**Title**
Tunde Csoban	Comparison of Predictive Modelling Methods for Predicition of Adverse Outcomes of pre-Eclampsia
Robin Muegge	Quantifying the Impact of National Lockdown on Covid-19 Deaths - Was it Really Worth it?
Nikola Ondrikova	Norovirus Reporting Patterns: Directions for Forecasting

Model Selection II	**Wednesday 28th July, 11:30-12:30**
Chair: Muhammad Mahmudul Hasan	
Speaker	**Title**
Getnet Melak Assegie	Multivariate Permutation Test for Big Data
Laura Freijeiro-González	Analysing the LASSO Covariates Selection Capability in the High-dimensional Framework under Dependence among Covariates
Lanxin Li	Bayesian Group LASSO Regression for Genome-Wide Association Studies

Medical Statistics	**Wednesday 28th July, 13:00-14:15**
Chair: Jinran Zhan	
Speaker	**Title**
Natalia Dygas	Analysing Women's Health During Pregnancy using Molecular Level "omics" Data
Pedro Fontes	Implications of Treatment Changes on Advanced Melanoma in Real-World Population
Samantha Ofili	Spatiotemporal Variability in Social, Emotional and Behavioural Development of Children
Xueqing Yin	Spatiotemporal Cluster Detection and Disease Risk Estimation using Clustering-based Adjacency Modelling

Probability	**Thursday 29th July, 9:45-10:45**
Chair: Qasem Tawhari	
Speaker	**Title**
Firdous Ahmad Mala	Probability of a Relation on a Set to be Transitive
Mariya Mamajiwala	Stochastic Dynamical Systems Developed on Riemannian Manifolds: Application to non-Convex Optimization
Laura Stewart	Shots in Boxes: A Simulation Study

Statistical Modelling I	**Thursday 29th July, 9:45-10:45**
Chair: Jordan Richards	
Speaker	**Title**
Eleanor D'Arcy	Accounting for Seasonality in Extreme Sea Level Estimation
Mahfuzur Rahman Khokan	Improving Power by Conditioning in Genetic Association Studies
Maeve Upton	General Additive Models for Relative Sea Level Change along the East Coast of North America

Keynote III	**Thursday 29th July, 11:00-12:00**
Chair: Tom Pinder	
Speaker	**Title**
Professor Mark Girolami	The Statistical Finite Element Method

Statistical Modelling II	**Thursday 29th July, 12:15-13:15**
Chair: Maeve Upton	
Speaker	**Title**
María Alonso-Pena	Contribution on Nonparametric Regression with Circular Variables
Qasem Tawhari	Random Spatial Graphs
Szymon Urbas	Modelling and Prediction of Clinical Trial Enrolment using a Time-inhomogeneous Hierarchical Approach

Abstracts

Contributions on Nonparametric Regression with Circular Variables

Date: 29th July
Day: Thursday
Time: 12:15-13:15
Room: Statistical Modelling II

María Alonso-Pena
Universidade de Santiago de Compostela

Many recent statistical works deal with estimation and inference methods for data supported on manifolds. A simple particular case is circular data, where the support of the variables is the unit circumference. Angles, directions or any periodic variable can be regarded as circular data. Examples of circular data are found on many different fields, such as biology (orientation and escape directions in animals), meteorology (wind and wave directions) or many others. The particular nature of this type of data must be taken into account when applying any statistical technique, and most estimation and inferential methods must be adapted to deal with the periodicity of the data. In particular, in this work we focus on the case of regression when one or all the variables are of a circular nature. We introduce new statistical techniques which allow studying the dependence between variables when at least one is circular.

Optimal Design of Clinical Trials in Multi-state Models

Date: 27th July
Day: Tuesday
Time: 14:15-15:15
Room: Clinical Trials

Hanan Alzahrani
King's College London University

A multi-state model is a stochastic process in which a point occupies one of several discrete states at anytime. The simplest multi-state model (two-state model) for survival data will be introduced in which the survival time in state 1 is recorded. We construct the exact D-optimal and Weighted A-optimal designs considering the Weibull model, which is a nonlinear model, as the distribution of the survival time (T). We also compute the same designs for the Exponential model as a special case of the Weibull. We optimize the design criterion numerically, using different prior values for the unknown parameters and N simulated data sets from the Weibull model, by an R program.

Date: 28th July
Day: Wednesday
Time: 11:30-12:30
Room: Model Selection II

Multivariate Permutation Test for Big Data

Getnet Melak Assegie[1], Stefano Bonnini[2]

[1]Parma University, Italy
[2]Ferrara University, Italy

BACKGROUND: In recent years, hypothesis testing of big datasets with a large number of response variables has been increasing. Parametric methods like Hotelling T-square test cannot be applied when the sample size is less than the number of outcomes. Moreover, homoscedasticity and normality assumptions are not often plausible. We propose a nonparametric solution based on a permutation test for multivariate-two sample problems.
METHOD: The combined permutation test(CPT) is powerful, flexible, and valid for multivariate responses with large number of components regardless of the sample sizes.
RESULTS: The simulations prove that CPT is more powerful than Hotelling T-square test and it is also a performant solution when the number of response variables is greater than sample sizes.
CONCLUSIONS: CPT is unbiased, powerful and consistent. CPT is a suitable solution for big data testing problems when sample size is lower than number of dependent variables.

Date: 27th July
Day: Tuesday
Time: 15:30-16:30
Room: Bayesian Statistics

Estimating the Proportion of Modern Contraceptive Methods supplied by the Public and Private Sectors using a Bayesian Hierarchical Penalized Spline Model

Hannah Comiskey[1], Dr. Niamh Cahill[1], Dr. Leontine Alkema[2]

[1]Hamilton Institute, Maynooth University
[2] School of Public Health & Health Sciences, University of Massachusetts Amherst

Nationally representative survey observations are considered to be reliable Family Planning indicators but are only available on average every 3-5 years. In order to bridge the data gaps, Track20 developed the “SS to EMU tool” to produce estimates of modern use (EMU) from routinely collected family planning service statistics. A key step in the tool involves accounting for potentially missing private sector data using the latest Demographic Heath Survey (DHS) estimates for the proportion of modern contraceptive methods supplied by each sector.

We propose a Bayesian hierarchical penalised spline model with multivariate normal spline coefficients, to account for across method correlations, to produce country specific annual estimates for the proportion of modern contraceptive methods coming from the public and private sectors. We provide estimates and uncertainty from 2000 to 2020 for 30 countries in the FP2030 initiative, who have both DHS data and Service Statistic data available after 2012.

Comparison of Predictive Modelling Methods for Prediction of Adverse Outcomes of Pre-eclampsia

Date: 28th July
Day: Wednesday
Time: 10:15-11:15
Room: Epidemiology

Tunde Csoban

University of Strathclyde

Pre-eclampsia is a pregnancy condition that can cause adverse maternal outcomes including death and serious morbidities. Current recommended models for prediction of adverse outcomes use logistic regression. The aim of this project is to explore alternative predictive modelling methods using multi-study data (n=7842).

Methods used were random forests, LASSO, ridge regression, Bayesian model averaging (BMA) and neural networks. Models were assessed on their ability to accurately classify patients into outcome- and risk-groups. The results showed good performance from all models, at least slightly improving upon the original model (area under ROC (AUC)=0.78, sensitivity (Se) =0.77; number predicted for low risk (n_{Low})=85, $n_{Moderate}$=1481, n_{high}=78). Random forests (AUC=0.82, Se=0.87; n_{Low} =554, $n_{Moderate}$=957, n_{high} =133) performed best, significantly improving on low-moderate-high classification. LASSO, Ridge regression and BMA performed similarly to the original model.

While machine learning does not offer a significant increase in AUC, our results demonstrate that three-group classification is improved.

Bayesian Variable Selection in Binary Quantile Regression Using a Quantile-specific Prior

Date: 27th July
Day: Tuesday
Time: 15:30-16:30
Room: Bayesian Statistics

Mai Dao

Texas Tech University

In this talk, we construct a Bayesian hierarchical modeling framework for simultaneously conducting parameter estimation and variable selection in binary quantile regression. We first impose the asymmetric Laplace distribution on the model errors and then specify a quantile-dependent prior for the regression coefficients, which allows researchers to set different priors for modeling different order of quantiles and thus yields great flexibility in Bayesian quantile modeling. By utilizing the normal-exponential mixture representation of the asymmetric Laplace distribution, we propose a novel three-stage computational scheme starting with an expectation-maximization algorithm and then a Gibbs sampler followed by an importance re-weighting step to draw independent Markov chain Monte Carlo samples from the full conditional posterior distributions of the unknown parameters. Simulation studies are conducted to compare the performance of the proposed Bayesian method with that of several existing ones in the literature. Finally, real-data applications are provided for illustrative purposes.

Date: 29th July
Day: Thursday
Time: 09:45-10:45
Room: Statistical Modelling I

Accounting for Seasonality in Extreme Sea Level Estimation

Eleanor D'Arcy
Lancaster University

Storm surges pose an increasing risk to coastline communities. These events, combined with high tide, can result in coastal flooding. To reduce the impact of storm surges, an accurate estimate of coastal flood risk is necessary. Specifically, estimates are required for the return level of sea levels (still water), which is the level with annual exceedance probability p. This estimate is used as an input to determine the height for a coastal defence, such as a sea wall. The return level estimation requires statistical analysis based on extreme value theory, as we need to know about the frequency of events that are more extreme than those previously observed.

Large storm surges exhibit seasonality, they are typically at their worst in the winter and least extreme in the summer. We focus on the skew surge: the difference between the observed and predicted high water within a tidal cycle. The seasonal pattern of skew surge differs from that of the tide, whose seasonality is driven astronomically, resulting in tidal peaks at the spring and autumn equinoxes. Hence, the worst levels of the two components of still water level are likely to peak at different times in the year, and so statistical methods that treat them as independent variables are likely to over-estimate return levels.

Our work aims to model how the distribution of skew surges changes over a year and we combine our results with the known seasonality of tides to derive estimates of still water level return levels.

Date: 28th July
Day: Wednesday
Time: 13:00-14:15
Room: Medical Statistics

Analysing Women's Health During Pregnancy using Molecular Level "omics" Data

Natalia Dygas
University of Strathclyde

Throughout pregnancy, the mother's body undergoes many physiological changes. The Yale Pregnancy Outcome Prediction Study looked at singleton pregnancies in women from a middle-class background. The aim of this project was to investigate the differences between inter/intra-placental metabolisms and identify the metabolites which showed significant differences between these metabolisms. Principal Component Analysis and Partial Least Squares-Discriminant Analysis were used to investigate the degrees of separation and similarity in inter/intra-placental sample groups. T-tests were also carried out to test the hypothesis that the means of the inter/intra-placental sample groups are equal, for each metabolite within the groups. Multivariate analyses showed that inter-placental samples had a higher degree of separation between groups, and similarity within the groups, than intra-placental samples. From the results, we can conclude that the metabolites which differ significantly in the inter-placental comparisons lead to a significant change in the metabolisms.

Implications of Treatment Changes on Advanced Melanoma in Real-world Population

Pedro Fontes

University of Strathclyde

Date: 28th July
Day: Wednesday
Time: 13:00-14:15
Room: Medical Statistics

The relative effectiveness of treatment paths remains unclear for melanoma, which caused over around 2700 deaths in the United Kingdom only and 60,712 deaths worldwide in 2018. Prior to 2010, dacarbazine was the standard treatment with patients presenting poor prognosis. Ongoing trials are investigating the possible synergism of combinational therapy. Since then, immunotherapy and targeted therapy have revolutionised treatment for advanced melanoma, significantly extending patients' lives. Observational studies can provide additional information. Patients may experience treatment changes. These changes affect patients survival and may cause adverse events. This retrospective study uses data on treatment outcomes of patients in the Greater Glasgow and Clyde region to explore the relative effectiveness of treatment pathways. Methodology follows survival analysis methods with Kaplan-Meier curves and Cox proportional-hazards models to estimate overall survival and influence of patient characteristics on survival, respectively. Findings from this research have shown that patients that switch treatments have longer survival that patients who do not switch.

Analysing the LASSO Covariates Selection Capability in the High-dimensional Framework under Dependence among Covariates

Laura Freijeiro-González

Universidade de Santiago de Compostela

Date: 28th July
Day: Wednesday
Time: 11:30-12:30
Room: Model Selection II

It is in the high-dimensional framework with $p > n$ where classic methods fail. Specifically, if we want to adjust a linear regression model we need to resort to penalization approaches as the well-known and widely employed LASSO regression. However, this methodology needs to satisfy some restrictive properties to guarantee optimality and it seems not to be the best option under dependence structures among covariates. We illustrate this behaviour by means of an extensive simulation study, making use of different dependence scenarios. Moreover, in order to search for improvements, we carry out a broad comparison with LASSO derivatives and alternatives. Eventually, we give some guidance about what procedures are the best in terms of the considered data nature.

Bayes Linear Emulation for a Complicated Computer Simulation

Muhammad Mahmudul Hasan

Durham University

Date: 28th July
Day: Wednesday
Time: 10:15-11:15
Room: Computational Statistics

The analysis of the output from a large-scale computer simulation experiment can pose a challenging problem in terms of size and computation. We consider output in the form of simulated crop yields from the Environmental Policy Integrated Climate (EPIC) model, which requires a large number of inputs - such as fertiliser levels, weather conditions, and crop rotations - inducing a high dimensional input space. We adopt a Bayes linear emulation rather than Bayesian approach to efficiently emulate crop yield as a function of the simulator fertiliser inputs. We explore different emulator diagnostics and present the results from emulation of a subset of the simulated EPIC data output via k-fold cross-validation and leave one out cross validation.

Date: 27th July
Day: Tuesday
Time: 14:15-15:15
Room: Clinical Trials

Assessing Patient Benefit with Semi-Markov Multi-state Models

A. Haris Jameel

University of Nottingham

In the context of oncological drug trials, competing risks models are used to model covariate dependence on time-to-event medical data. In particular, the semi-parametric Cox proportional hazards model is traditionally used to establish treatment efficacy based on patient response to treatment. However, an effective treatment may not be one that benefits the patient as such drugs often have high rates of patient drop-out. Such drop-outs are detrimental from the perspective of patients, wasting their already limited remaining lifespan. Semi-Markov multi-state models are well known and can be used to model the event history of a patient in a drug trial. Comparing the time spent in each state between different groups of patients can offer additional information besides the efficacy of the drug, thus allowing for an alternative set of tools to quantify whether patients can benefit from a potential life-prolonging treatment. Patients, payers, and caretakers can then use this information to make better-informed decisions about such treatments.

Date: 27th July
Day: Tuesday
Time: 15:30-16:30
Room: Bayesian
Statistics

Decision Support for Large Offshore Wind Farms via Bayesian Optimisation

Jack Kennedy

Newcastle University

Large offshore wind farms require corrective maintenance to keep availability (performance) to a satisfactory level. This requires the wind farm operator to buy and store specialist spare components, which are difficult to get hold of, at a nearby warehouse. This poses two questions: (1) how many spare parts of different types should the operator buy to ensure high availability? and (2) which warehouse should the operator store the components in? We investigate the consequences of such decisions via a complex, stochastic offshore wind farm simulator known as Athena coupled with a utility function which quantifies how preferable any set of consequences are. In a Bayesian decision analysis, the optimal decision is the decision which maximises the decision maker's expected utility. Since Athena is expensive and stochastic, we utilise Bayesian Optimisation to find a 'good' solution to the decision problem in a computationally efficient manner.

Improving Power by Conditioning in Genetic Association Studies

Date: 29th July
Day: Thursday
Time: 09:45-10:45
Room: Statistical Modelling I

Mahfuzur Rahman Khokan
Open University

In the area of statistical genetics, classical genome-wide association studies (GWAS) assess the association between a biological characteristic and genetic variants, working with one variant at a time in a regression model, and reporting the most significant associations. However, in many cases there are known databases of major genetic variants that have a substantial effect on the trait. In such situations it makes sense statistically to condition on these major variants to improve power in detecting associations with new variants; but this is not a common practice in GWAS applications.

In this study, we show theoretically and computationally how conducting a joint analysis of the genetic variants in a multivariate regression model, where the estimated effect of a new variant is conditioned upon some major variants, can improve the performance of the model in terms of reducing the standard error and improving the power. The amount of gain of power will depend on the correlation between the response and the covariates, as well as correlation between the covariates. We further show that conditional results can sometimes be obtained from publicly available summary statistics reported for univariate associations in published GWAS studies, even when the individual-level data are unavailable.

A prominent example of such a trait is skin colour, for which there are many studies consistently identifying a handful of major genes. We will look into a dataset for over 6,500 mixed ethnicity Latin Americans to see how the conditioning process can improve the detection power of GWAS studies and identify new genetic variants in such a situation.

Bayesian Group LASSO Regression for Genome-wide Association Studies

Date: 28th July
Day: Wednesday
Time: 11:30-12:30
Room: Model Selection II

Lanxin Li
University of Glasgow

Genome-wide association studies (GWAS) are experiments aimed at searching the genome for genetic changes (SNPs) that contribute to a trait of interest, by looking at hundreds of thousands of SNPs simultaneously. This can be represented as a high-dimensional variable selection problem, where SNPs are treated as candidate predictors. However, it is difficult for most existing variable selection methods to accurately detect SNPs underlying complex diseases (like heart disease), due to weak association signals from SNPs, between-SNP correlations (due to linkage disequilibrium), and the sheer imbalance between the sizes of the sample and predictor sets. We propose to use the Bayesian group Lasso framework, which makes use of the correlation structure between SNPs and reduces the computational cost with large SNP datasets through grouping, along with a powerful population MCMC method to sample from the posterior distribution, to improve the accuracy and efficiency of SNP detection in GWAS.

Date: 29th July
Day: Thursday
Time: 09:45-10:45
Room: Probability

Probability of a Relation on a Set to be Transitive

Firdous Ahmad Mala
Govt. Degree College Sopore

The study of probability and statistics demands reasonable expertise in combinatorics. In fact, any problem with enumeration is potentially a problem of probability. In this paper, the probability of a randomly chosen relation on a set to be one among the various special classes of relations is discussed. In particular, lower bounds for the probability of a relation to be a transitive relation are discussed using certain new recursions.

Date: 29th July
Day: Thursday
Time: 09:45-10:45
Room: Probability

Stochastic Dynamical Systems Developed on Riemannian Manifolds: Application to Non-convex Optimization

Mariya Mamajiwala
University College London

We propose a method for developing the flows of stochastic dynamical systems, posed as Ito's stochastic differential equations, on a Riemannian manifold identified through a suitably constructed metric. The framework used for the stochastic development, viz. an orthonormal frame bundle that relates a vector on the tangent space of the manifold to its counterpart in the Euclidean space of the same dimension, is the same as that used for developing a standard Brownian motion on the manifold. Mainly drawing upon some aspects of the energetics so as to constrain the flow according to any known or prescribed conditions, we show how to expediently arrive at a suitable Riemannian metric and the associated connection. We demonstrate the application of the method to a few benchmark problems in non-convex optimization. The simplicity of the method and the sharp contrast in its performance vis-á-vis the correspondent Euclidean schemes in our numerical work provide a compelling evidence to its potential.

Date: 27th July
Day: Tuesday
Time: 14:15-15:15
Room: Model Selection I

Information-criteria-based Model Selection for Neural Networks

Andrew McInerney
University of Limerick

Neural network model selection is usually carried out using a 'trial-and-error' approach, with varying initial weights and network architecture. However, the calculation of an associated likelihood function opens the door to information-criteria-based model and variable selection, and likelihood-based confidence intervals for network weights. Novel 'bottom-up' and 'top-down' model selection methods are proposed using the Bayesian information criterion for feedforward multi-layer perceptrons whereby the optimal weights for one model are carried over to the next. Compared to the standard trial-and-error search through the space of models, this is both more computationally efficient and has an increased probability of recovering the true model. Simulation studies are used to evaluate the performance of the proposed methods, and an application on real data is investigated.

Penalized Multi-Parameter Regression Modelling

Date: 27th July
Day: Tuesday
Time: 14:15-15:15
Room: Model Selection I

Laura McQuaid
University of Limerick

Multi-parameter regression (MPR) modelling refers to the approach whereby covariates enter a parametric model through multiple distributional parameters simultaneously (e.g., scale and shape parameters), allowing more complex covariate effects to be captured. On the other hand, penalized estimation procedures such as the least absolute shrinkage and selection operator (LASSO) and adaptive LASSO are commonly used to perform continuous variable selection - but they have primarily been developed for classical regression problems where the covariates enter only through a single distributional parameter. Therefore, we develop a penalized MPR modelling framework and investigate its performance through simulation studies and real data analysis. We consider the application area of survival analysis, but the methodology can equally be applied to other areas.

Quantifying the Impact of National Lockdown on Covid-19 Deaths – Was it Really Worth it?

Date: 28th July
Day: Wednesday
Time: 10:15-11:15
Room: Epidemiology

Robin Muegge
University of Glasgow

This talk will discuss the analysis of weekly Covid-19 data from England, spanning from the beginning of the pandemic (March 2020) to the most recent weeks (early June 2021). The focus is on the number of deaths due to Covid-19, and on the impact that restrictions issued by the government may have on these counts. Emphasis will be given on the differences by local authority (the areal unit) over time (using week as temporal unit), under consideration of local restrictions. The goal is to estimate the impact of local restrictions issued by the government on the count of deaths, to get a better understanding of the effectiveness of different measures, regarding the degree of restrictions and the time the measures were put in place. Ideally, the findings will be useful in giving advice on how to regulate the spread of the virus effectively in the future.

Multivariate Extremes for Nuclear Regulation

Date: 28th July
Day: Wednesday
Time:16:30-17:45
Room: π-minute Theses

Callum Murphy-Barltrop
Lancaster University

Often, statistics are produced that summarise the 'body' of the data (where the majority of values lie), such as the mean. However, sometimes it's more important to consider the 'tail' of the data; these are the values that deviate significantly from the body of the data. We label such events as 'extremes' and we model these values using extreme value theory.

In this talk, we introduce a statistic that is often used to summarise multivariate extremal behaviour, which we term a return curve. These curves are defined, for a given probability p, to be all values of a random vector for which the joint survival probability is equal to p. For applications where the risk from combinations of two (or more) variables is considered important, such as the analysis of coastal structures, these curves may allow resources to be better allocated than if the extremes of the variables were considered separately.

We also introduce novel uncertainty characterisation tools for return curve estimation and compare methods for curve estimation. Furthermore, we demonstrate our approach using a case study and illustrate some of the potential applications of return curves.

Date: 28th July
Day: Wednesday
Time: 13:00-14:15
Room: Medical Statistics

Spatiotemporal Variability in Social, Emotional and Behavioural Development of Children

Samantha Ofili
University of Strathclyde

The residential area influences where, when, and for whom contextual differences in developmental outcomes occur. Multilevel models that combine individual, temporal and spatial data can be developed to identify the children and areas most vulnerable to poor outcomes. This requires the model to accurately specify the variation between individuals, time periods and areas using random effects. The Child Mental Health in Education (CHiME) study has social, emotional and behavioural outcomes for children aged 4-10 in Glasgow between 2010 and 2017 linked to demographic and geographic information. Using candidate multilevel Bayesian models identified from the literature, this project aims to evaluate how well the model specifications (including computation, distribution, parameters and priors) describe the CHiME data. Further work will address how these models can be developed to explain more complex structures of the data (including spatiotemporal interactions and residential mobility) using simulated data.

Date: 28th July
Day: Wednesday
Time: 10:15-11:15
Room: Epidemiology

Norovirus Reporting Patterns: Directions for Forecasting

Nikola Ondrikova
University of Liverpool

Norovirus has a higher level of under-reporting in England compared to other intestinal infectious agents such as *Campylobacter* or *Salmonella*, despite being recognised as the most common cause of gastroenteritis globally. We investigated heterogeneity in passive surveillance system to improve understanding of differences in reporting and laboratory testing practices of norovirus in England. Multiple model formulations were compared, and the best performing model was determined by proper scoring rules based on one-week-ahead predictions. The best performing model highlighted atypically large and small amounts of reporting by comparison with the average in England. Endemic random intercept varied from the lowest in East Midlands in those in the under 5 year age-group (0.36, CI 0.18–0.72) to the highest in the same age group in South West (3.00, CI 1.68–5.35). Our findings provide guidance for development of norovirus forecasting tools.

Smooth BIC Variable Selection Procedure for Heteroscedastic Data

Date: 27th July
Day: Tuesday
Time: 14:15-15:15
Room: Model Selection I

Meadhbh O'Neill
University of Limerick

Modern variable selection procedures revolve around penalization methods to execute simultaneous model selection and estimation. A popular method is the lasso (least absolute shrinkage and selection operator), which contains a tuning parameter. This parameter is typically tuned by minimising the cross-validation error or Bayesian information criterion (BIC) but this can be computationally intensive as it involves fitting an array of different models and selecting the best one. However, we have developed a procedure based on the so-called "smooth BIC" in which the tuning parameter is automatically selected. We also extend this model selection procedure to the so-called "multi-parameter regression" framework, which is more flexible than classical regression modelling. Multi-parameter regression introduces flexibility by taking account of the effect of covariates through multiple distributional parameters simultaneously., e.g. mean and variance. These models are useful in the context of normal linear regression when the process under study exhibits heteroscedastic behaviour.

How to Increase the Sample Size: The Use of Rolling Windows

Date: 28th July
Day: Wednesday
Time: 10:15-11:15
Room: Computational Statistics

Virgilio Pérez
University of Valencia

One of the key concepts to consider when conducting a research is sample size because as more and more information become available, the uncertainty reduces. In this study we have focused on how to extend the sample size without the need to obtain new data, and without barely increasing the associated error levels. To solve the problem of insufficient information, we propose to pool data, i.e. to admit that close subjects behave similar. To verify this assumption, we used a large database with more than 700,000 observations. We have found that there are indeed certain time variables that evolve very smoothly, such as political ideology, so it is not unreasonable to pool data to increase the sample size. The proposed method describes the use of the rolling windows, achieving the described objective of increasing the sample size without barely increasing the estimation error.

A Review of Simulation Studies Comparing Statistical and Machine Learning Approaches to Risk Prediction for Time-to-Event Data

Date: 28th July
Day: Wednesday
Time: 10:15-11:15
Room: Computational Statistics

Hayley Smith
University of Leicester

In risk prediction modelling, there is increasing interest in the application of the machine learning approach to time-to-event data and comparing these methods to common statistical approaches, usually the standard Cox model. However, the simulation settings and choice of comparison methods can be considered very limited, and arguably biased towards machine learning approaches. Ensuring comparisons are unbiased and comprehensive can be difficult to implement in practice (1), however, may be achievable with Austin's (2) approach, where every model forms the data-generating mechanism for each scenario. We conducted a methodological review of simulation studies that compare machine learning and statistical methods for risk prediction to identify where comparisons may be lacking, biased towards one approach or misleading. A limited number of articles were identified and key information was often missing or unclear. In this talk, I will discuss the results of the review and offer recommendations for the issues raised.

References:
1. Boulesteix A-L, Wilson R, Hapfelmeier A. Towards evidence-based computational statistics: lessons from clinical research on the role and design of real-data benchmark studies. BMC Medical Research Methodology. 2017;17(1).
2. Austin PC, Harrell FE, Steyerberg EW. Predictive performance of machine and statistical learning methods: Impact of data-generating processes on external validity in the "large N, small p" setting. Statistical Methods in Medical Research. 2021:096228022110028.

Date: 29th July
Day: Thursday
Time: 09:45-10:45
Room: Probability

Shots in Boxes: A Simulation Study

Laura Stewart
University of Glasgow

Suppose we allocate n shots to N boxes using the uniform distribution. It has been shown that under certain conditions on n and N the distribution of the number of occupied boxes is either Poisson or Normal. Now, suppose we instead want to have multiple layers of N boxes. In the first layer we will throw $n = N$ shots. For the remaining layers the number of shots is given by the number of occupied boxes in the previous layer. Also, if two or more shots land in the same box then they merge so the number of occupied boxes is non-increasing as we move through the layers. This multi-layer scheme can be used to model processes involving objects merging over time. We can simulate this process to gain insight into results that we can then try to prove for the number of occupied boxes for different layers and for different N.

Date: 29th July
Day: Thursday
Time: 12:15-13:15
Room: Statistical Modelling II

Random Spatial Graphs

Qasem Tawhari
Durham University

We look at a class of random spatial graphs constructed on the points of a Poisson point process in the unit square, with edges defined by a geometrical rule based on proximity. Specifically, each point is joined by an edge to its nearest neighbour in a given direction specified by a cone. The unrestricted case is the ordinary nearest-neighbour graph; the restricted case is a version of the minimal directed spanning forest (MDSF) introduced by Bhatt & Roy. These graphs have been widely used for modelling networks with spatial content, such as in the communications sector, social networks, and transportation networks. The large-sample asymptotic behaviour of the total edge length of these graphs is our main interest. For the ordinary nearest-neighbour graph, the appropriate central limit theorem is due to Avram & Bertsimas. For the MDSF, the limit theory is known (Penrose & Wade) in two special cases, namely the 'south' and 'south-west' versions: here the limit is not normal, due to the presence of long edges near to the boundary.

In this talk, I will describe ongoing work to extend the limit theory to the case of general cones; depending on the parameters, the limit distribution may be normal, or the convolution of a normal distribution with a non-normal element due to boundary effects. This is joint work with Nicholas Georgiou and Andrew Wade.

General Additive Models for Relative Sea Level Change along the East Coast of North America

Date: 29th July
Day: Thursday
Time: 09:45-10:45
Room: Statistical Modelling I

Maeve Upton
Maynooth University

Relative sea level is the observed sea level change with respect to the Earth and forms part of a complex pattern of interactions, varying in time and space. The aim of this project is to develop statistical models to examine historic sea level changes for North America's and Ireland's Atlantic Coast using proxy data. The statistical approach employed is that of extensions of General Additive Models, which allow separate components of sea level to be modelled individually and for smooth rates of change and accelerations to be calculated. The model is built in a Bayesian framework which uses prior information to capture the evolution of sea level change. The proxy data is collected as salt-marsh sediment cores and dated using biological and geochemical sea level indicators. By combining statistical models and proxy data, results have shown that current sea level along North America's east coast is the highest it has been in at least the last 15 centuries.

Modelling and Prediction of Clinical Trial Enrolment using a Time-inhomogeneous Hierarchical Approach

Date: 29th July
Day: Thursday
Time: 12:15-13:15
Room: Statistical Modelling II

Szymon Urbas
Lancaster University

Clinical trials are the standard in medical treatment development. They are composed of lengthy and often very costly phases composed of different types of studies. During Phases II and III, the studies often experience delays due to insufficient enrolment of subjects. In the talk, I will discuss our recent work in improving the industry standard for modelling and predicting the enrolment of patients. The work centres around a time-inhomogeneous hierarchical model which models the enrolment patterns at site-level. Additionally, I will discuss some practical considerations when implementing novel statistical models.

Bias induced during the estimation of quality-adjusted life-years

Date: 28th July
Day: Wednesday
Time:16:30-17:45
Room: π-minute Theses

Alexandra Welsh
Lancaster University

Quality-adjusted life-years (QALYs) are a summary measure used to evaluate the effectiveness of medical treatments in terms of both quality and length of life. One method used to estimate QALYs is the area under the time-utility curve (AUC). However, this approach may induce bias, due to its inability to capture the dependency between the quality of life measures and the survival time. A simulation study is conducted to assess the bias induced when estimating QALYs using the AUC method, using data including censored individuals and missing responses. A further aim of the project is to determine the suitability of joint longitudinal-survival models, using both standard and "reverse" timescales, for the estimation of QALYs, in order to reduce the potential bias and increase efficiency.

Date: 28th July
Day: Wednesday
Time: 13:00-14:15
Room: Medical Statistics

Spatiotemporal Cluster Detection and Disease Risk Estimation using Clustering-based Adjacency Modelling

Xueqing Yin
University of Glasgow

Globally spatially smooth conditional autoregressive (CAR) models are typically used to capture the spatial autocorrelation in areal unit disease count data when estimating the spatio-temporal trends in disease risk. In these models the spatial autocorrelation structure is typically induced by a binary neighbourhood matrix based on a border sharing specification, such that spatial correlation is always enforced between geographically neighbouring areas. However, enforcing such correlation in the model will mask any discontinuities in the disease risk surface, thus impeding the detection of clusters of areas that exhibit higher or lower risks compared to their neighbours. Therefore, we propose novel methodology to account for these discontinuities via a two-stage modelling approach, which either forces the spatial clusters to be the same for all time periods or allows them to evolve dynamically over time.

Date: 27th July
Day: Tuesday
Time: 14:15-15:15
Room: Clinical Trials

Comparison of Frequentist and Bayesian Methods for Borrowing Historical Two-arm Data in Clinical Trials

Jinran Zhan
University of Warwick

Incorporating historical data into a current trial is a potential strategy to reduce development costs and patient burden in clinical trials. However, this can lead to the inflation of the type I error rate. Therefore, it is important to choose an appropriate method to ensure that the amount of strength borrowed from the historical trial is adjusted to the agreement between the current and previous studies. Several approaches have been proposed for integrating historical control data, but there is relatively little research on borrowing from both arms of a single historical two-arm trial. We compare frequentist and Bayesian approaches in terms of their operating characteristics and explore factors influencing the borrowing behaviour in the two-arm setting.

Sponsorship Information

Royal Statistical Society

Founded in 1834, the Royal Statistical Society (RSS) is one of the world's leading organisations for promoting the importance of statistics and data. The organisation advocates for the key role of statistics and data in society, and works to ensure that policy formulation and decision making are informed by evidence for the public good. The RSS is also a professional body for statisticians and data analysts in the UK and around the world, with more than 10,000 members to date.

The Lancashire and East Cumbria RSS local group organises an annual program of meetings for statisticians (and non-statisticians) in the Lancaster and Preston area to discuss statistical methods and their applications. Aligning with one of their key priorities to support early career statisticians, the local group is funding four prizes for the conference.

STOR-i Centre for Doctoral Training

The STOR-i Centre for Doctoral Training (CDT) is a joint partnership between the Department of Mathematics and Statistics and the Department of Management Science at Lancaster University. The primary aim of the centre is to offer a four-year doctoral training programme in Statistics and Operational Research, with a specific focus on solving problems for industry. STOR-i is funded by the Engineering and Physical Sciences Research Council, which funds CDTs in various disciplines across the country.

45th RSC Nottingham

The next conference (RSC2022) will be hosted in Notthingham – stay tuned!