UNSW SYDNEY | Australia's Global University

School of Mathematics and Statistics

Honours in Statistics Handbook

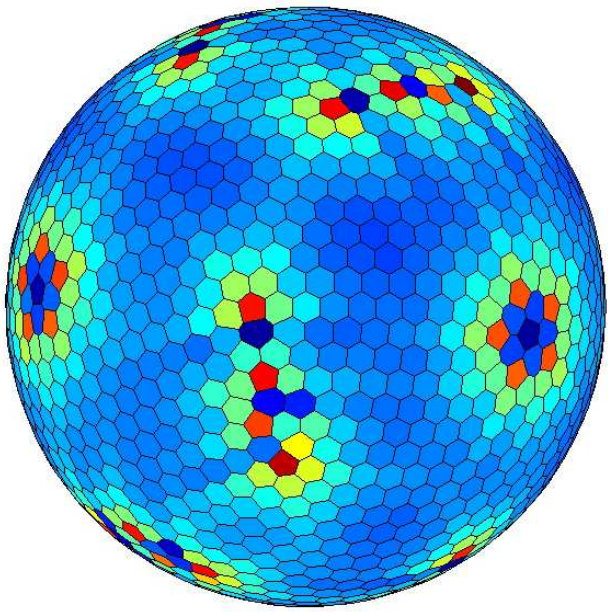
Contents

Overview

The School of Mathematics and Statistics at the University of New South Wales, Sydney, provides an exciting venue for undergraduate Honours program, both for Australian and international students. Situated close to the heart and beaches of Australia's largest city, we have become one of the most successful mathematics schools in the country. The Department of Statistics is one of the largest in Australia.

The Honours year introduces students to the investigative and research aspects of knowledge and consists of advanced lecture courses, an Honours thesis and seminar participation. We offer expert supervision across a wide range of areas in statistics and financial mathematics. Our Honours students are supervised in heir Honours project by some of Australia's finest Statisticians and Mathematicians.

Over the past decade, the School of Mathematics and Statistics has produced many Honours graduates. These highly trained graduates have been snapped up by a variety of employers, often before their Honours year is finished. Many of them have entered the financial sector, finding well-paid jobs with firms such as Macquarie Bank, Citibank and AMP. Others have entered government employment, or other areas including the electricity and the telecommunication industries. Others again have continued on to higher degrees, such as PhD studies, in Australia or overseas. Employers value the problem-solving, report-writing and research skills induced by this training.



Entry requirements

For UNSW students

To enter Honours in Statistics, students must have

- Completed a Statistics major in the Science and/or Mathematics program, including at least 30 units of credit in Level III Mathematics
- An average above 70% in their Level III statistics major courses (MATH3901, MATH3911, MATH3821 and at least one of MATH3831, MATH3841, MATH3851)
- An average above 70% in their Level III Mathematics courses

With the permission of the Head of School (or Honours Co-ordinator) a student may be allowed into Honours without having satisfied the specific requirements.

In order that you have sufficient background to attempt the courses in the Honours year, it is recommended that you discuss your selection of Level III courses with the Honours Coordinator or another academic adviser.

Honours students enrol full-time or part-time in one of the Honours programs MATH4903/MATH4904.

For students outside UNSW

Students who are interested in transferring to UNSW to complete or undertake their Honours year should consult the Honours coordinator as soon as possible to ascertain whether this is feasible. Because our Honours courses are based around the background knowledge obtained in the first three years of the UNSW degree, even students with excellent 3 year degrees from elsewhere sometimes lack the background to attempt a full quota of courses. In such cases we often recommend that students complete an extra semester at UNSW before starting Honours in order to fill in some of the missing prerequisites.

Honours Rules

Requirements

The Honours year will be 48uoc. The compulsory thesis will be 18uoc. In addition to the thesis work each student will undertake 5 courses of 6uoc each.

Honours students each present a seminar of 30 minutes on their thesis to members of the mathematics staff, interested visitors and other students. Students should also attend any appropriate seminars in their thesis area. This requirement enables students to become familiar with the range of current research in Mathematics and to see how to present a research seminar.

Courses

An honours student must undertake at least 24uoc from the Department of Statistics, where the thesis counts for 18uoc.

External Courses: With the permission of the relevant Honours Coordinator, a student may take courses from other disciplines at UNSW, other mathematics schools (for example, at University of Sydney), and external courses such as those taught at the AMSI Summer School.

Choosing your courses

Honours year students in statistics should take a broad range of courses in ALL AREAS of statistics. You are encouraged to take one course offered by Pure or Applied mathematics departments. Students should discuss their course selection with the honours coordinator prior to the commencement of each semester.

Seminar practice

At the end of their first session or the beginning of their second session, each student is required to present their thesis work in the same style as their year-end oral presentation.

The thesis

Students will write an honours thesis on their honours topic. The thesis normally includes a literature survey and a critical analysis of the topic area. This should prepare you for the problem-solving and report-writing aspects of future employment, or for progression to a research degree. Each student works under the supervision of one or more members of the Department on an investigation of some topic in Statistics that is currently an area of active research. Prospective students should start talking to staff members about possible thesis topics well before they start their Honours year. An early decision about a topic will facilitate an early start on reading. Supervision by individual staff members is dependent on staff agreement and availability. The thesis will be assessed for quality in four major areas (see below), each of which is important. The mark for the thesis will be made up of 90% for the written thesis and 10% for an oral presentation. The written thesis will be assessed by two markers other than the supervisor(s), and each marker will provide a written assessment and grade(s) based on the following:

- Exposition: Clarity of the presentation. Sufficient introductory and summary material. Organisation and style of the presentation.
- Literature coverage: Adequate coverage of related material in the field. Placing the topic in a wider context.
- Critical analysis and insight: Understanding of the problem and/or model. Quality of the discussion. Discussion of the advantages and limitations of the

problem/method.

- Originality: E.g. by modifying or extending earlier theory or methods, or by developing new examples, or by an application to a new area.

Oral presentation

Typically in week 13 of the last semester of their honours year, students each present a seminar of 30 minutes on their thesis to members of the Statistics staff, interested visitors and other students. This presentation is worth 10% of their thesis mark. The presentation will be assessed on: Knowledge displayed; motivation presented for the study of the topic; description of contributions/achievements; description of results; clarity of verbal discussion; clarity of slides/figures; keeping to time; and responses to questions.

Honours grade

Sum of thesis grade (worth 37.5%) and 5 courses (worth 12.5% each), possibly scaled.

Time line

The following is a general guide to how work on your thesis should progress. If you think that a major variation is warranted, please discuss this with either your supervisor or the Honours Coordinator.

- Select supervisor and topic - Before the start of your Honours year
- Research, reading, discussion, understanding - First semester mostly
- Outline of thesis and significant piece of writing - By the beginning of your last semester
- Give a substantial draft to supervisor - End of week 8 of your last semester
- Talk - Last teaching week of your last semester
- Final submission - by 5pm on the Friday of the last teaching week in the last (typically the 2nd) thesis semester

Thesis due date

The honours thesis is due at 5pm on the final day of week 13 in the final semester of honours candidature. If the thesis is late with no good reason, the final thesis mark f will be calculated as

$$f = \begin{cases} r & \text{if } r < 50 \\ 50 + (r - 50)e^{-0.03n} & \text{if } r \geq 50 \end{cases}$$

where r is the recommended mark (before taking lateness into account), and n is the number of days that the thesis is overdue. In the case of illness or other extenuating

circumstances, the late penalty will be determined by agreement between the two thesis assessors and the honours coordinator.

Staff Directory

Title	Name	Room	Phone	Email
Head of Department:				
Prof	Scott Sisson	1034	9385 7027	Scott.Sisson@unsw.edu.au
Statistics Honours Coordinator:				
Dr	Feng Chen	1031	9385 7026	Feng.Chen@unsw.edu.au
Director of Honours Programs:				
Prof.	Gary Froyland	3060	9385 7050	g.froyland@unsw.edu.au
Director of Undergraduate Studies:				
Dr	John Steele	5103	9385 7060	J.Steele@unsw.edu.au
Student Services Manager:				
Ms	Julie Hebblewhite	3088	9385 7053	J.Hebblewhite@unsw.edu.au
Academic staff:				
Dr	Zdravko Botev	2056	9385 7475	botev@unsw.edu.au
Dr	Leung Chan	1036	9385 7021	leung.chan@unsw.edu.au
Dr	Diana Combe	1032	9385 7022	diana@unsw.edu.au
Prof	Pierre Del Moral			p.del-moral@unsw.edu.au
Prof	William Dunsmuir	2057	9385 7035	W.Dunsmuir@unsw.edu.au
Dr	Yanan Fan	2055	9385 7034	Y.Fan@unsw.edu.au
Dr	Gery Geenens	2053	9385 7032	ggeenens@unsw.edu.au
Prof	Ben Goldys			b.goldys@unsw.edu.au
Dr	Pierre Lafaye de Micheaux	2050	9385 7029	lafaye@unsw.edu.au
Dr	Libo Li	1035	9385 7025	libo.li@unsw.edu.au
A/Prof	Jake Olivier	2051	9385 6656	j.olivier@unsw.edu.au
A/Prof	Spiridon Penev	1038	9385 7023	S.Penev@unsw.edu.au
Dr	Donna Salopek	2054	9385 7030	dm.salopek@unsw.edu.au
Dr	Jakub Stoklosa	3071	9385 4723	j.stoklosa@unsw.edu.au
Dr	Peter Straka	1033	9385 7024	p.straka@unsw.edu.au
Prof	David Warton	2052	9385 7031	David.Warton@unsw.edu.au

Honours projects

Examples of possible honours projects are given in the following pages, for each staff member.

Dr Zdravko Botev

Exact sampling from Bayesian Bridge model. One of the simplest Bayesian models for measurements $y_1, \ldots, y_m$ is the linear regression with model parameter $(\boldsymbol{\beta}, \sigma)$, likelihood $\boldsymbol{Y}|(\boldsymbol{\beta}, \sigma) \sim \mathsf{N}(\mathrm{X}\boldsymbol{\beta}, \sigma^2\mathrm{I})$, and prior density $p(\boldsymbol{\beta}, \sigma)$ of the special form:

$$p(\boldsymbol{\beta}, \sigma) \propto \exp\left(-\sum_i \lambda_i |\beta_i/\sigma|^\alpha\right), \quad 0 < \alpha \leq 1,\ \lambda_i \geq 0.$$

This prior give rise to the *Bayesian Bridge* model, which is of great interest in applications, because it may allow us to gain more reliable information about the model parameters, $\boldsymbol{\beta}$ and σ, than it is possible with least squares estimation. Unfortunately, as yet, simulation from the posterior is only possible via *approximate* Markov chain Monte Carlo sampling.

In this project you will explore a new computational method for *exact* Monte Carlo sampling from the Bayesian posterior probability density function

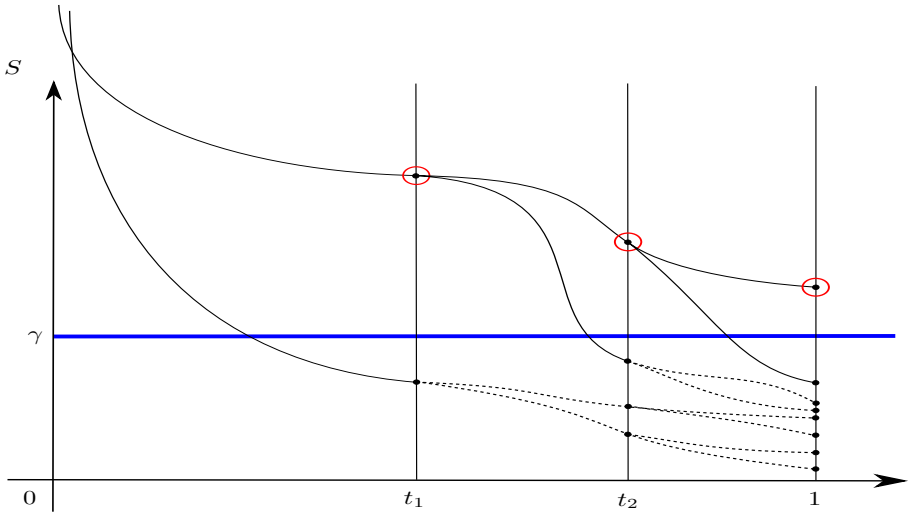
$$p(\boldsymbol{\beta}, \sigma|\boldsymbol{y}) \propto \exp\left(-\frac{\|\boldsymbol{y}-\mathrm{X}\boldsymbol{\beta}\|^2}{2\sigma^2} - m\ln\sigma - \sum_i \lambda_i |\beta_i/\sigma|^\alpha\right)$$

You will test the new computational technique on large datasets and draw appropriate inferences.

Monte Carlo Splitting method for quasi-monotone models.

Consider the estimation of the small probability $\ell = \mathbb{P}(S(\boldsymbol{X}) \geq \gamma)$, where $\boldsymbol{X}$ is drawn from a d-dimensional pdf f, the threshold γ is large enough to make ℓ very small, and S is a quasi-monotone function (e.g., $x_i \leq y_i,\ i = 1, \ldots, d$ implies $S(\boldsymbol{x}) \leq S(\boldsymbol{y})$). Such problems frequently arise in portfolio credit risk assessment, where estimating ℓ is of interest, because it is the probability of bankruptcy or default.

In its original form, the Monte Carlo splitting method for simulation of Markov processes cannot be applied to estimate ℓ, because there is no underlying continuous time-dependent Markov process that can be split (see Figure).



However, it is possible to embed the density f within a continuous time Markov process, making is possible to apply the splitting method after all. In this project you will explore the theoretical properties of this special Monte Carlo splitting algorithm with the goal of making the algorithm more computationally efficient. This project may involve collaboration with researchers in Rennes in France and Montreal in Canada.

Dr Leung Chan

My research interests examine how to use mathematical techniques to solve some practical problems in finance. Modern techniques from stochastic processes, statistics, partial differential equations and numerical analysis are now used in mathematical finance. My research interests will focus on the following areas:

The first topic of interest is to investigate the application of regime switching to finance. Regime switching is known to model business and economic cycles. I wish to apply regime switching to exotic options such as barrier options, Asian options, passport options, etc.

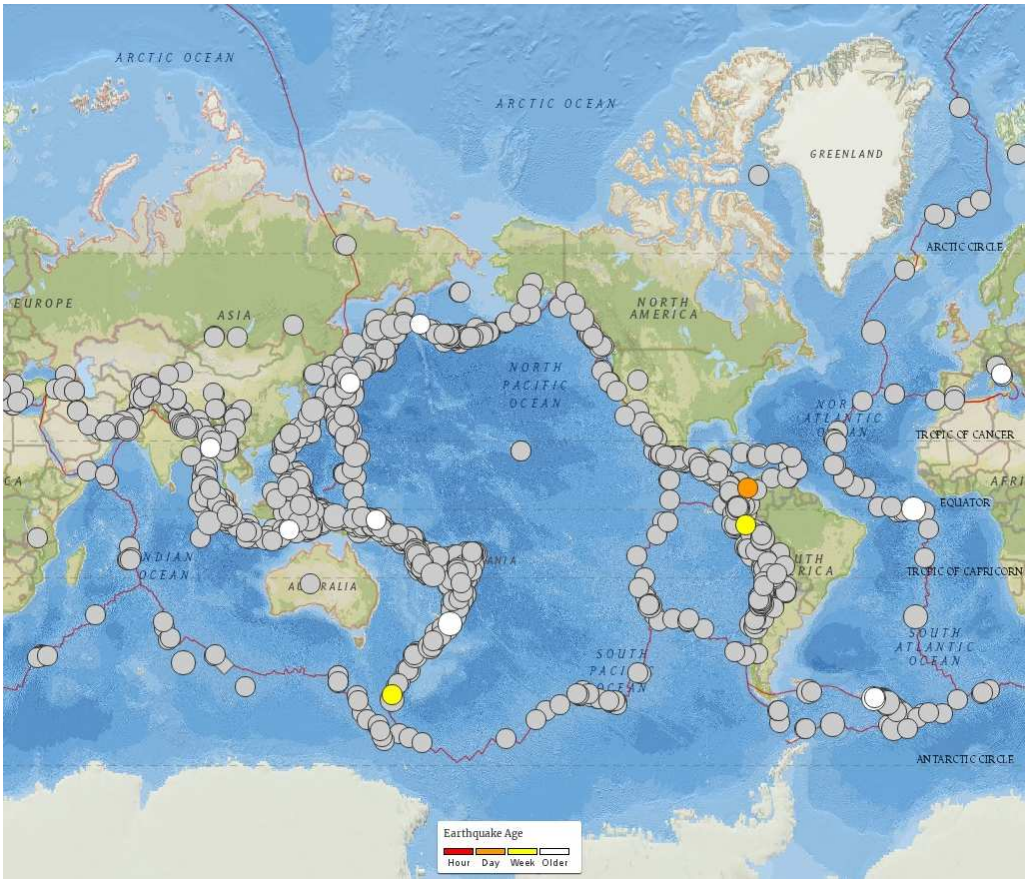
The second topic of interest is to solve high dimensional problems arising in mathematical finance. The Monte Carlo simulation is the best method to deal with high dimensional problems so far as I know. I am interested in developing a novel approach which is better than the Monte Carlo simulation.

The third interest is to investigate a numerical solution of the stochastic differential equations (SDEs). The multiple stochastic integrals arisen from higher order strong approximations or weak approximations are unstable for the current available numerical methods. I am interested in investigating a novel approach which can provide a stable numerical method of SDEs with higher order convergence.

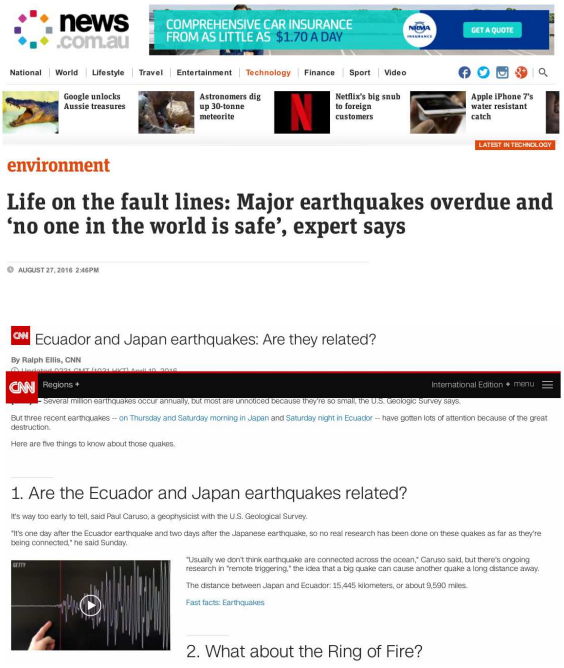
Finally, a fourth interest is to investigate some problems which arise from the volatility derivative market. In particular we shall investigate volatility derivatives such as variance swaps, volatility swaps, options on the realized variance, etc.

Dr Feng Chen

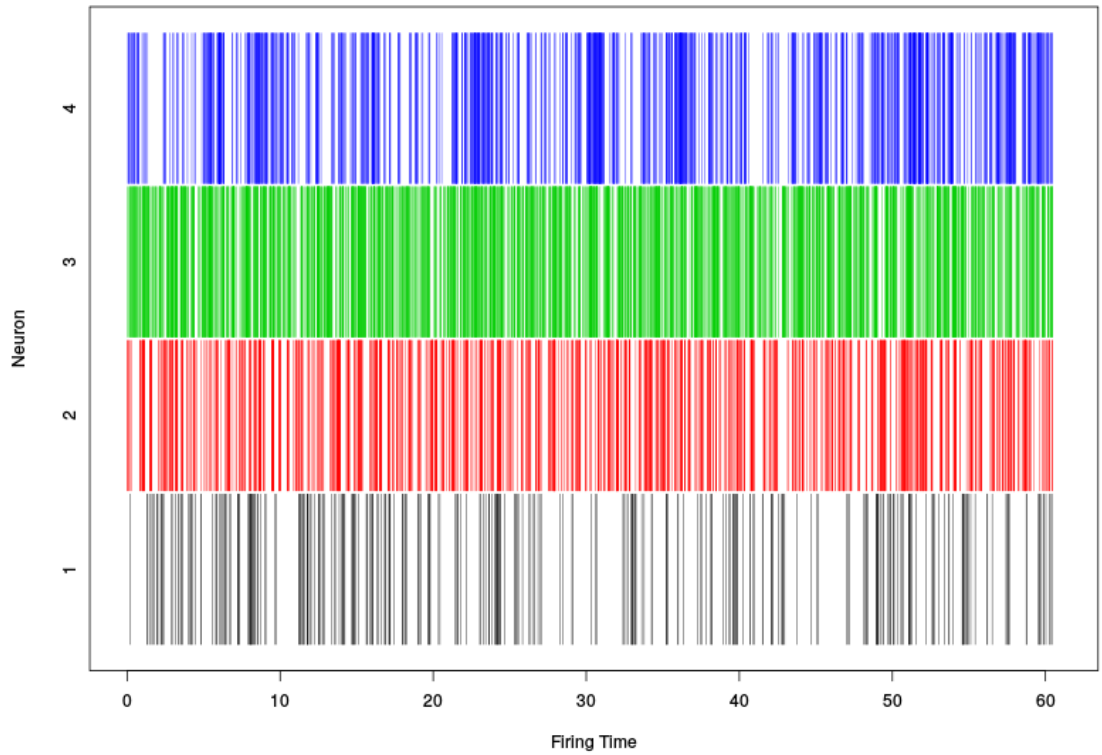
Project 1: Modelling and prediction of major earthquakes



In every 5 minutes, an earthquake occurs somewhere in the world, and a major earthquake (Mg 6+) hits somewhere in every 3.4 days. By modelling the sequence of earthquakes in a region of interest, this project helps us to answer questions like: "Is a major earthquake overdue in the region?" (cf. the news.com.au headline story), "Is the recent earthquake related to a certain quake that happened previously at a certain time?" (cf. the CNN headline story), "what is the probability of having a major earthquake in the next month?".



Project 2: Neural spike train analysis



Neurons can propagate signals rapidly over large distances by generating characteristic electrical pulses called action potentials: voltage spikes that can travel down nerve fibers. If the brief duration of an action potential (about 1 ms) is ignored, an action potential sequence can be characterized simply by a series of all-or-none point events in time, and is referred to as a spike train. By modelling the apparently random temporal patterns of the firing of the voltage spikes by a group of neurons, using e.g. point processes, we can infer about the self- and mutual-excitation relationships among the neurons, which then helps us understand the functional dependence among the neurons.

Dr Diana Combe

All my research is in areas of combinatorics and algebra. I am interested in patterns and symmetries of patterns: both in investigating the existence and number of patterns with particular properties or symmetries (or automorphisms), and also in exploiting the symmetries of patterns to construct combinatorial objects of various kinds.

The specific areas in which I work are combinatorial designs and related graphs, graph theory and graph labellings. Much of my work is theoretical, but some is computational and involves large computer searches. A project could be theoretical or involve a large amount of computing.

Some suggested areas for projects:

A Generalized Bhaskar Rao design over a finite group G is a rectangular array of entries which are either 0 or group element entries, and which satisfy particular properties (of orthogonality, which I won't describe here). From these can be constructed group divisible designs with high levels of symmetries, which can be used in experimental design. One issue is that of existence - for what parameters there do exist generalized Bhaskar Rao designs. these investigations involve designs and graphsand groups: and the geometry of one can be used to gain insights to the others.

Examples of projects in this area would be: (i) Give an explicit proof of the necessary and sufficient conditions for a generalized Bhaskar Rao design with block size 3 over the cyclic group of order 8. (ii) Determine the structure of the graph associated with generalized Bhaskar Rao designs with a particular set of parameters. (iii) Explore the practical applications of generalized Bhaskar Rao designs: for example in determining new group divisible designs. (This could be a survey and bibliography.)

Graph labellings can be labellings of vertices (or of edges, or of both edges and vertices) over the natural numbers, (or over the integers, or over finite or infinite groups) satisfying particular constraints. Graph labellings can be used to solve particular practical or theoretical problems or to explore the structure and geometry of the graph or the labelling set (or group).

Some suggested projects in this area would be: (i) Determine the edge-magic (or vertex-magic, or both edge-magic and vertex-magic) labellings for a family of graphs over finite abelian groups. (ii) Explore extending a type of graph labelling to a labelling of directed graphs (and, possibly investigate the practical applications).

Professor Pierre Del Moral

My major research interest is in designing and analyzing stochastic models and methods for nonlinear estimation and optimization problems. I am particularly interested in advanced Monte Carlo methodologies such as sequential/adaptive Monte Carlo methods including mean field type interacting particle models. In the last decades, the theoretical and the numerical aspects of this interdisciplinary research field have been under rapid development. These modern stochastic techniques have become one of the most active contact points between probability theory, Bayesian inference, statistical machine learning, information theory, theoretical chemistry and quantum physics, financial mathematics, signal processing, risk analysis, and several other domains in engineering and computer sciences.

Some suggested areas for projects involving some open mathematical research problems:

Project 1: Branching approximations of Feynman-Kac formulae. This project is concerned with designing particle simulation techniques based on adaptive branching rules. The current project is investigating the numerical and the mathematical aspects of some branching algorithms recently developed in computational physics for solving ground state molecular energies. The application domains considered in this project are not limited to computational physics. In this project, we apply these branching models to rare event simulation and nonlinear filtering problems. This project will involve collaborations the department of physics and quantum chemistry of the University Paul Sabatier, Toulouse, France.

Project 2: The Ensemble Kalman Filter is a prediction algorithm currently used in meteorological forecasting and data assimilation problem. This stochastic algorithm can be interpreted in terms of McKean-Vlasov type particle models. This project is concerned with the numerical and the theoretical analysis of this class of model, including the exponential concentration analysis and the long time behavior of this class of interacting particle model. This project will involve collaborations the INRIA Research Center (the French national institute for research in computer science and control) in Rennes, France.

Professor William Dunsmuir

My major research interest is in developing models and methods for estimation in time series of correlated discrete valued observations arising particularly in public health policy evaluations and financial time series. Analysis of discrete valued time series present substantial challenges to computation, methodology and theory of inference. The field is still rather underdeveloped compared to the analysis of continuous valued time series. Examples arise in modelling the impacts of environmental exposures and the impacts of policy changes on the spatio-temporal distribution of disease and accident data. Analysis of these impacts often requires modelling of binary or small count data collected in small areas on short time scales. Methods that can handle large collections of such data are under development. In high frequency financial transactions data very large data sets are available for analysis in order to assess various models of market microstructure.

Project 1: Regression modelling in multiple time series of counts. This project is concerned with developing models and estimation methods for multiple time series of counts in which there is interest in testing if impacts of regressors are the same or not in subsets of the series. All approaches are based on the generalized linear autoregressive moving average models for exponential family responses for which Professor Dunsmuir has developed theory, methods and software.

The current project is investigating the inclusion of random effects into the serial dependence models. Applications are to examining the impact of lowering the legal blood alcohol level in drivers on vehicle fatalities across geographic regions and to investigating the impact of lowering and raising the minimum legal drinking age on suicide rates in teenagers across US states.

Project 2: Modelling high frequency financial transactions data. This project involves developing models for the durations between transactions as well as the amount by which the price changes at each trade. The primary investigation is concerned with assessing the stability of the models and parameters specifying them from day to day and developing hierarchical models for the the parameter change process.

Dr Yanan Fan

I am a Bayesian statistician who is primarily interested in the development of Monte Carlo simulation techniques, which is vital for Bayesian inference.

Bayesian inference is about the incorporation of prior information into a likelihood-based inference. In such situations, inference on the posterior distribution (prior and likelihood) can be numerically challenging, as they often involve high dimensional integrals.

One of my primary interest is in the development of Monte Carlo methods for numerical integration. Among these methods, I am particularly interested in Markov chain Monte Carlo and transdimensional Markov Chain Monte Carlo, sequential importance sampling and sequential Monte Carlo.

For example, Markov chain Monte Carlo (MCMC) methods rely on the construction of a Markov chain to sample from the posterior distribution. These methods rely heavily upon the construction of a proposal distribution (which should be similar to the posterior distribution). A recent interest in Markov chain Monte Carlo is how to use the information collected from the Markov chain sample path to guide the construction of the proposal distribution.

A potential project may be to consider several approaches recently developed in the literature, which makes use of a mixture of proposal distributions and/or running several Markov chains simultaneously. Another potential project may be to consider a particular use of MCMC in Bayesian variable selection problems, where there may be thousands of variables to consider.

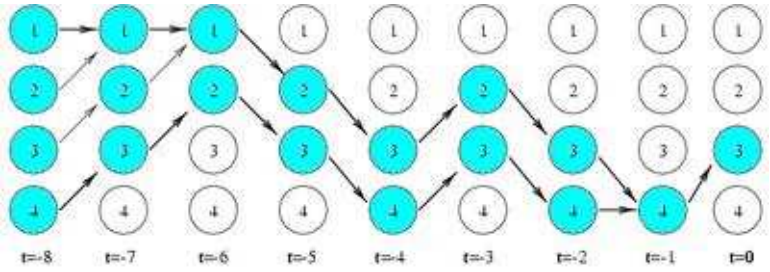


Figure 1: Coupling of Markov chains: four Markov chains starting from different states are made to couple.

I also have a number of potential projects co-supervised with Dr Don Weatherburn (Director of NSW Bureau of Crime Statistics and research). Below are some examples of possible topics.

How accurately can we forecast adult/juvenile prison population size from information about police contacts with suspected offenders?

Correctional administrators need the earliest possible warning of increases in custodial populations. It is therefore of interest to know whether and to what extent we

can forecast juvenile and adult custody populations from information about police contacts with suspected offenders. A possible complication here is that there may be endogeneity between police contacts and custodial populations.

What effect does school retention have on juvenile offending?

There has been much debate about whether policies which encourage school retention help reduce juvenile crime (by keeping kids out of trouble and/or improving their employment prospects). Our earlier study on this issue was inconclusive because we did not have much variation in the outcome measure (burglary) and because we did not specifically measure the number of juveniles apprehended by police.

What is the size of the population of amphetamine users and is it increasing?

Arrests for amphetamine type substance (ATS) use have been increasing but we have no idea how many regular ATS users there are. Several methods have been developed for estimating the size of hidden populations. We have substantial data on the arrest history of amphetamine users. A question of interest is whether it is possible to estimate the size of the ATS user population.

How sensitive are our measures of re-offending?

We spend a lot of time and money evaluating the effect of various programs on re-offending. Usually we compare rates of reconviction in a treatment and control group, making adjustments for various confounders. Only a small fraction of offences, however, result in a conviction. We have often wondered whether it would be possible to simulate policy effects of varying sizes (Monte Carlo methods?) and see how big they have to be before we can detect them.

What effect does pro-active policing have on crime?

Police use a number of methods to prevent and control crime. These include moving on suspected offenders, searching suspected offences, checking compliance with bail conditions and writing intelligence reports on the activity of suspected offenders. No-one in Australia has examined the effect of this kind of proactive policing on crime. Possible panel analysis.

Dr Gery Geenens

Most of my research lies in developing nonparametric and semiparametric methods in various contexts. Traditional parametric models assume that the functional form of the statistical objects of interest (distribution of a random variable, regression function, etc.) is exactly known up to a finite number or parameters. This is restrictive and, in case of model misspecification, can lead to erroneous conclusions. In contrast, nonparametric models keep the structural prior assumptions as weak as possible, really 'letting the data speak for themselves' (as it is commonly quoted). Usually, this (almost) total flexibility is reached by using statistical models that are infinite-dimensional. This, of course, brings in some new challenges, both in theory and in practice.

More specifically, several topics I am currently interested in are the following.

Nonparametric copula modelling: Copula modelling has emerged as a major research area of statistics. A bivariate copula function C is the joint cumulative distribution function (cdf) of a bivariate random vector whose marginals are Uniform over $[0,1]$, i.e., $C : \mathcal{I} \doteq [0,1]^2 \to [0,1] : (u,v) \to C(u,v) = \mathbb{P}(U \leq u, V \leq v)$, where $U \sim \mathcal{U}_{[0,1]}$, $V \sim \mathcal{U}_{[0,1]}$. Copulas arise naturally as a mere consequence of two well-known facts. First, the *probability-integral transform* result: if $X \sim F_X$ is continuous, then $F_X(X) \sim \mathcal{U}_{[0,1]}$; and second, *Sklar's theorem*: for any continuous bivariate cdf is F_{XY}, there exists a unique function C such that

$$F_{XY}(x,y) = C(F_X(x), F_Y(y)) \qquad \forall (x,y) \in \mathbb{R}^2, \tag{$\star$}$$

where F_X and F_Y are the marginals of F_{XY}. From the above, this function C is, indeed, a copula, called the copula of F_{XY}. It describes how F_X and F_Y 'interact' to produce the joint F_{XY} and clearly disjoints the marginal behaviours of X and Y from their dependence structure, hence the attractiveness of the approach. Today, copulas are used extensively in statistical modelling in all areas, from quantitative finance and insurance to medicine and climatology. Therefore, empirically estimating a copula function from a bivariate sample $\{(X_i, Y_i)\}_{i=1}^n$ drawn from F_{XY} has become an important problem of modern statistical modelling.

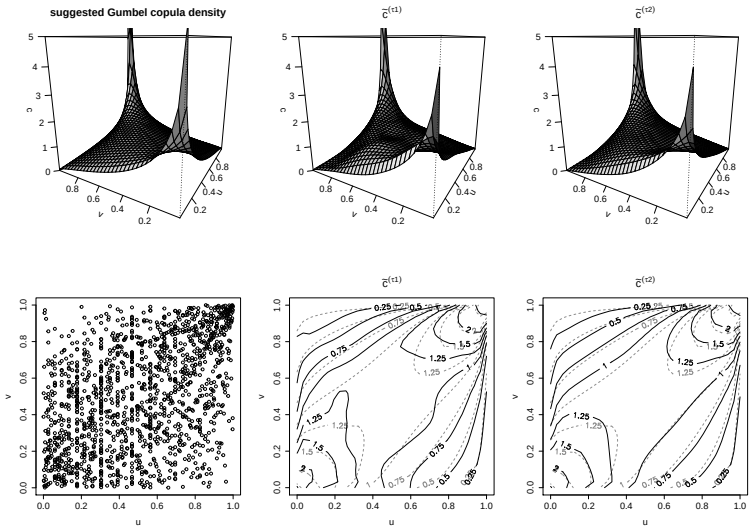
One can differentiate each side of ($\star$) to obtain

$$f_{XY}(x,y) = f_X(x) f_Y(y) c(F_X(x), F_Y(y)), \tag{$\star\star$}$$

where $f_{XY}(x,y)$ is the usual joint probability density of (X,Y) with marginal densities f_X and f_Y, and

$$c(u,v) = \frac{\partial^2 C}{\partial u \partial v}(u,v)$$

is the so-called copula density. Nonparametrically estimating this copula density is *very* challenging.



Nonparametric copula-based conditional density estimation: More than a regression model $m(x) = \mathbb{E}(Y|X = x)$, the conditional density $f_{Y|X}(y|x)$ provides complete information about the relationship between the dependent variable Y and the regressor X. The conditional density of Y given X, i.e.

$$f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)}, \qquad (\star\star\star)$$

has, from $(\star\star)$, an alternative representation in terms of the copula density:

$$f_{Y|X}(y|x) = f_Y(y)c(F_X(x), F_Y(y)).$$

A nonparametricestimator of $f_{Y|X}(y|x)$ based on this representation has multiple practical advantages over the more classical estimators of conditional densities where both the numerator and the denominator of $(\star\star\star)$ are estimated separately. It, however, requires nonparametric estimation of the copula density.

Nonparametric high-dimensional density estimation via pair-copula construction: Equation $(\star\star)$ can be generalised to more than two dimensions. For three random variables X_1, X_2, X_3, for instance, the joint density f_{123} can be written

$$f_{123}(x_1,x_2,x_3) = f_1(x_1)f_2(x_2)f_3(x_3)c_{12}(F_1(x_1),F_2(x_2))c_{13}(F_1(x_1),F_3(x_3)) \\ \times c_{23|1}(F_{2|1}(x_2|x_1),F_{3|1}(x_3|x_1)).$$

In general, any d-variate joint density can be expressed as a product of its d marginals times $\binom{d}{2} = d(d-1)/2$ pair-copula densities, acting on several different conditional probability distributions. This 'pair-copula construction' breaks down the high-dimensional density in a product of lower-dimensional objects, that should be easier to estimate. This offers a promising path for nonparametrically estimating any high dimensional probability density, which is a difficult problem. Again, efficient estimation of copula densities is required to achieve this.

Other related applied studies are also considered. So far, most of the applied studies lies within a parametric framework, so that any use of nonparametric methods in applications might be innovative. For instance, I have recently developed a nonparametric model aiming at analysing football (soccer) results, which shows interesting and up to now ignored patterns. All suggestions are welcome.

Professor Ben Goldys (University of Sydney and UNSW)

Stochastic, ordinary and partial, differential equations are indispensable tools for the modelling and analysis of complex phenomena evolving randomly in space and time. They are important for modelling of climate, the evolution of prices on the stock market, changes in distribution of species, the theory of magnetic materials, Quantum Field Theory and General Relativity. A fascinating feature of the modern theory of stochastic differential equations is breaking the boundaries between Statistics, Pure and Applied Mathematics. In particular, stochastic differential equations have become a tool to tackle certain difficult problems in Differential Geometry, Analysis, analysis and the theory of deterministic partial differential equations. Two areas of potential projects are listed below and others can be discussed directly with me.

1. Geometric measure theory in infinite dimensions
2. Probabilistic approach to harmonic maps and applications

For more detailed description of projects, please follow this link.

Dr Pierre Lafaye de Micheaux

Project description to be included. Please contact Dr Lafaye de Micheaux directly.

Dr Libo Li

The general theory of stochastic processes was developed by the French school of probabilists in the 1970's. In the recent years, this theory has found increasing applications in mathematical finance. Broadly speaking, I am interested in the study of the general theory of stochastic processes and its applications to financial mathematics. The areas which I am currently working on are:

- The general theory of stochastic processes, in particular, the theory of enlargement of filtration.
- Study of random times, specially pseudo-stopping times and their related properties.

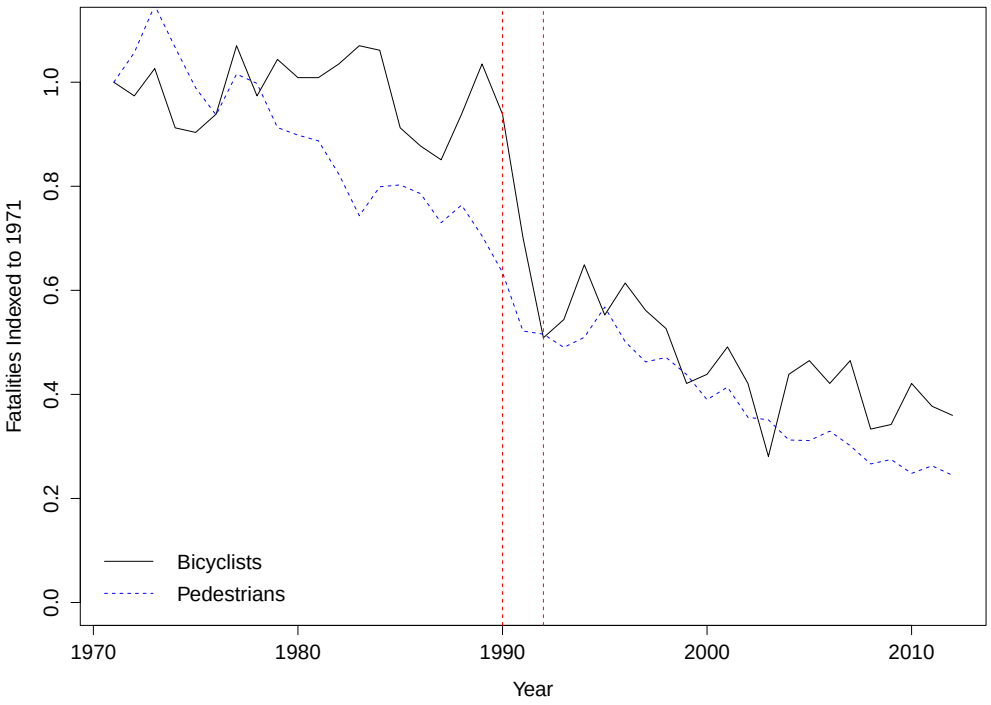
I am also interested in the *parametrix method* : This method can find its origins in the study of PDEs, however, it can be used to obtain an asymptotic expansion of density to stochastic differential equation, in particular, Lévy driven stochastic differential equation. This method can lead to a stochastic representation of the density, which can be computed using Monte Carlo simulation.

Associate Professor Jake Olivier

My research focus involves statistical methodology for the analysis of epidemiological and population health data. I regularly collaborate with the Transport and Road Safety (TARS) research centre and researchers within the Faculty of Medicine.

Current projects include:

Assessing Population Interventions: There are many statistical challenges in quasi-experimental designs due to lack of randomisation. This makes it difficult to assess the impact of population-based interventions due to unmeasured confounding. Some examples of interest are assessing the effects of bicycle helmet legislation and measures to limit access to guns. Interrupted time series, with or without a control, offer an improvement over other methods. However, little work has been done in this area and there is little standardization of analytic approaches across various fields such that varying analyses can lead to disparate conclusions using the same data source.



Effect Sizes: It is well known that statistical significance is a function of sample size. This is problematic because important effects can go unnoticed in studies with small sample sizes, and unimportant effects can give small p-values from large studies. Effect sizes can complement significance testing by removing the influence of sample size. A lot of work on this topic can be found in psychological research but little has been done on epidemiological measures like the odds ratio, relative risk or the hazards ratio.

Associate Professor Spiridon Penev

There are several areas in which my research interests are focused.

1. Wavelet methods in non-parametric inference. These include applications in density estimation, non-parametric regression and in signal analysis. Wavelets are orthonormal bases used for series expansions of curves that exhibit local irregularities. When estimating spatially inhomogeneous curves wavelets outperform traditional nonparametric methods. Typical wavelet estimators are of non-linear threshold-type. My research is focused on improving their flexibility by making a finer balance between stochastic approximation terms in the mean integrated squared error. I also make data-based recommendations for choosing tuning parameters in the estimation. My recent interest in wavelet methods is in applying them for shape-constrained estimation, for analyzing non-regularly sampled images and surfaces, and in applying Bayesian inference techniques in wavelet methodology.

2. Saddlepoint approximations for densities and tail-area probabilities of certain statistics turn out to be surprisingly accurate down to small sample sizes. Although being asymptotic in spirit, they sometimes give accurate approximations even down to a sample size of one. I have derived the saddlepoint approximation of the joint density of slope and intercept in the Linear Structural Relationship model. Similar in spirit to Saddlepoint is the Wiener germ approximation which I have applied to the non-central chi-square distribution and its quantiles. It performs better than any other approximations of the non-central chi-square in the literature.

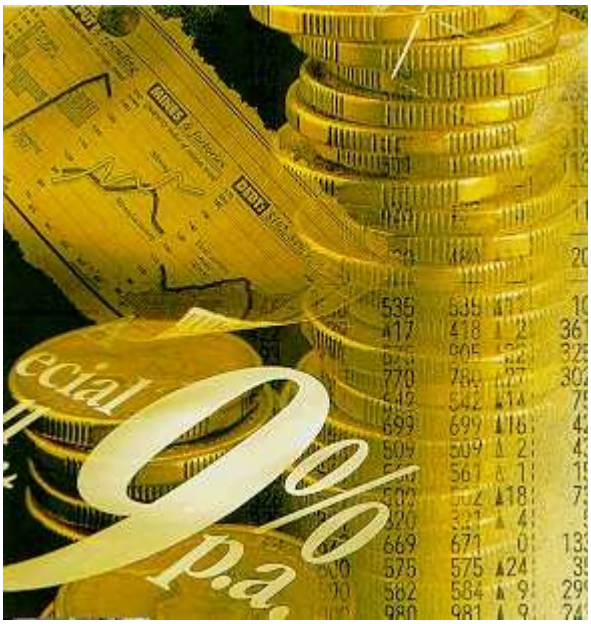
   (Higher order) Edgeworth expansions deliver better approximations to the limiting distribution of a statistic in comparison to the Central Limit Theorem. They may be non-trivial to derive for complicated estimators such as the kernel estimator of the p-th quantile. We study the Edgeworth expansion of the latter and demonstrate the improvement in comparison to the normal approximation.

3. Latent Variable Models (LVM) is an area of significant interest to me. I have worked on aspects of model choice and model equivalence in LVM. These are practically relevant issues- model equivalence appears more often in LVM than in other statistical models and represents a threat to the validity of inferences. I am also interested in inference for latent correlations and for scale reliability. I also have studied the relationship between the concepts of maximal reliability and maximal validity and have established useful inequalities between them. Within the setting of the LVM, I am interested in a subset of models called GLLAMM (generalized linear latent and mixed models). They are general enough yet with a sufficient structure to allow flexible modeling of responses of mixed type.

4. Investigation of dependencies and associations. Studying dependence of random variables when their joint distribution is not multivariate normal is important. A single correlation coefficient is not enough to describe the dependence in such cases and copula functions are a useful tool. A particularly simple copula is the Archimedean copula. We use spline methods in estimating Archimedean copulas.

Dr Donna Salopek

My research interests are financial modeling in general. In particular I am interested in non-semimartingales (eg. fractional Brownian motion) modeling, Levy modeling, and pricing of a variety of financial products. I am also interested in general problems in stochastic processes and stochastic analysis. For instance, I am currently working on stochastic evolution equations that are driven by Hcylindrical fractional Brownian motion and its possible application to financial mathematics.



Professor Scott Sisson

I am interested in Bayesian and computational statistics, Monte Carlo methods, symbolic data analysis, likelihood-free inferential methods, and extreme value theory, and the application and development of these in a broad range of application areas, including big data for health research, climate extremes, ecology, population genetics etc.

Projects are available in all of the below areas. For further information, including research preprints, please visit: http://www.maths.unsw.edu.au/~scott

- **Bayesian methods:** formulate both parameter and model uncertainty in distributional form, naturally incorporating these uncertainties into statistical inference. In practice, inference is performed by numerical sampling from this distribution. This is often extremely challenging, as the distribution may exhibit numerous complex features.

- **Monte Carlo methods:** Stochastic simulation algorithms commonly used to simulate from posterior distributions used for this task include rejection and importance sampling, Markov chain Monte Carlo (MCMC), trans-dimensional MCMC and sequential Monte Carlo.

- **"Likelihood-free" statistical inference:** Within classical and Bayesian inference, the ability to numerically evaluate the likelihood function is a critical requirement. However, suppose that you wish to consider models which are sufficiently realistic, that the likelihood function is computationally intractable, which causes inferential problems. "Likelihood-free" methods focus on the development and application of classical and Bayesian inferential procedures in this setting. These include approximate Bayesian computation (ABC) and pseudo-marginal MCMC methods.

- **Symbolic data analysis:** Sometimes data are generated in non-standard forms such as random-intervals/hypercubes, histograms, weighted lists or general distribution functions (mostly due to measurement error). Symbolic data analysis is a technique to analyse data where each datapoint is a distribution. This can be quite useful when analysing big and complex datasets, as these data can by collapsed down to a few thousand symbols (resulting in a much smaller dataset), and then analysed directly.

- **Environmental and climate extremes:** Statistical inference is typically interested in the mean levels of a process. But what happens if you're interested in extreme levels? This is important in the study of e.g. extreme rainfall, temperature, wind speeds and droughts. The limiting distribution of a standardised sample maxima is a Generalised extreme value distribution. This distribution may be used to model, e.g., the maximum daily temperature in a

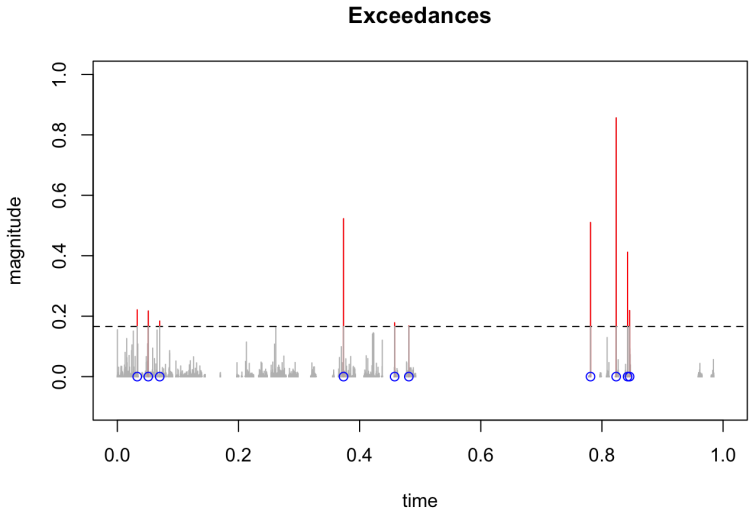
year, and then provide predictions on, say, 1 in 100-year temperature events. Other asymptotic extreme value theory results provide a suite of potential frameworks for modelling extreme process levels.

Dr Jakub Stoklosa

Project description to be included. Please contact Dr Stoklosa directly for more information.

Dr Peter Straka

Many empirically observed sequences exhibit infinite mean inter-arrival times, i.e. times between observations are drawn from a heavy-tailed distribution. This means that events occur in bursts, with periods of heavy activity alternating with quiet periods. In many situations magnitudes are attached to events, e.g. for earthquakes, neuron voltages and trade volumes, and the interest is in the prediction of the time and size of the *largest* event within a given time interval. The Continuous Time Random Maxima process (CTRM), or max-renewal process, models the dynamics of these types of extremes.



Project 1: Statistical Inference for CTRM.

Limit theorems are known for the distributions of occurrence times and occurrence magnitudes, where the limit is taken with respect to the number of observations going to infinity. This project is developing statistical inference for the model parameters of CTRMs. A main challenge lies in the intractability of Mittag-Leffler distributions, which calls for the development of computational models (in R, Python, Mathematica or Matlab), and another challenge in the modelling of parameter uncertainty, since the distributional assumptions do not hold exactly.

Project 2: Testing independence of waiting times and magnitudes.

Limit theorems (for large numbers of waiting times and observations) are available which characterize the joint distribution of waiting time and event magnitude. But before a model for dependence (which is more complex) should be considered, the independence of the two should be statistically tested. This project will develop independence tests for heavy-tailed distributions.

Professor David Warton

David Warton leads the Eco-Stats Research Group, a team of ecological statistics researchers and students working to improve methods for making use of data to answer research questions commonly asked in ecology. Our research ranges from theory and methods research (particularly regarding the analysis of high dimensional data) to applied research (introducing modern methods of analysis to ecology). We collaborate with ecologists and statisticians in Australia and internationally, in projects which have received over $2M in funding from the Australian Research Council.

Current research focuses on two distinct topics: (1) Developing model-based approaches to studying ecological communities and their responses to environmental impacts such as climate change; and (2) Novel methods for modelling the distribution of species as a function of environmental variables. There is considerable demand in ecology for modern statistical methods, and considerable potential in terms of developing new methodology inspired by ecological problems. Examples of current Eco-Stats research:

- Penalised likelihood estimation of point process models (with applications predicting species distribution)
- Fast variable selection for correlated count data
- Finite mixture of regression models to cluster species by environmental response

$5,000 Scholarships available, contact David for more information and some project ideas, see http://www.eco-stats.unsw.edu.au

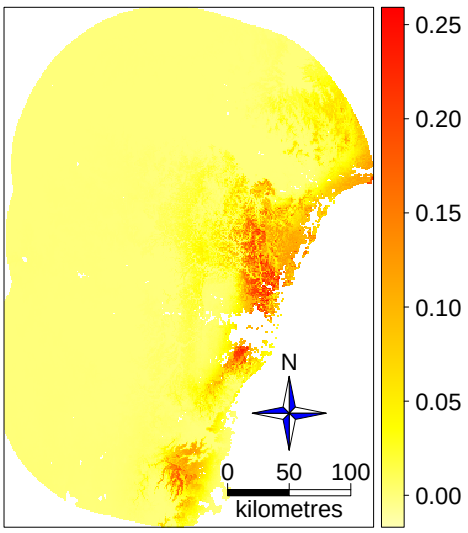


Figure 2: A model for the distribution of Sydney Red Gum in the Blue Mountains region, as a function of climatic variables. We are currently using models like this one to predict the potential effects of climate change on biodiversity.

Projects proposed by industry partners

The following are Westpac sponsored projects, you will be given the opportunity to work on a real-world problem, co-operate with industry professionals, and you will have access to actual data from Westpac.

- **Predicting Credit Card Defaults with High Frequency Transactional Data.**

  Credit scoring models or "scorecards" are widely used in the retail banking sector to assess default likelihood for loans. Whilst a broad range of factors are typically considered during the build of a default likelihood model, the most commonly used factors tend to be a set of highly summarised month end transactional data elements, such as balance at month end or credit utilisation at month end. Until recently, costs and computational limitations have acted as a barrier inhibiting unrestrained exploration of higher frequency transactional data for the derivation of new data elements potentially useful for improving default likelihood model discriminatory power and therefore commercial value.

  The objective of this project is to explore of a number of dimensionality reduction and advanced variable selection techniques (e.g. genetic algorithms) applied to high frequency transactional data, with the goal of optimizing and improving probability of default models.

- **Credit Decision Optimization**

  Mathematical decision optimization can yield a competitive advantage in many situations. We will perform a numerical investigation into the application of decision making algorithms to a particular problem in finance. The objective will be one that is widely sought after, to maximise expected profit. The question will be whether to approve a loan to a potential customer, and the decision will be a simple yes or no answer, based on data about the applicant.

  We will select two or three known optimization algorithms, possibly modifying the algorithm to better suit the problem, and compare the optimal strategies that they produce. There will be scope to choose different algorithms based on suitability and interest, some candidates are: the simplex algorithm (well understood, a good model for comparison), iterative methods such as gradient descent, or heuristic methods such as genetic algorithms or particle swarm optimization.

- **Modelling Probability of Default for Bank Loans**

  This project is focused on investigating a range of parametric and semi-parametric modelling techniques applied to probability of default estimation for lending portfolios.

The challenge of probability of default estimation is broad and lends itself to application of a wide range of alternative empirical modelling techniques open to investigation through this project, including:

  - Application of parametric and semi-parametric survival models with time varying covariates. These methods have potential application in a range of areas including estimating multi-year probability of default, constructing time to default probability densities and modelling cure probabilities for already defaulted loans.
  - Modelling credit rating migrations and default transition probabilities with Markov chains or other discrete time stochastic models.
  - Modelling effects of changes in credit underwriting standards and credit quality on default probabilities using time-series modelling techniques such as GARCH, ARIMA and Vector-ARIMA.

- **Modelling Exposure at Default for Bank Loans** Banks accredited by their regulators to use the Advanced Internal Ratings Based (A-IRB) approach are required to provide their own estimates (relying on statistical and analytical models) for calculating their minimum credit capital for bank loans that they issue. These estimates are outlined in the Basel Accord and comprise of the following three credit risk components:

  - Probability of Default (PD) The probability a customer will fail to make full and timely repayments and default on their loan.
  - Exposure At Default (EAD) The expected value of the loan balance at the time of default.
  - Loss Given Default (LGD) The amount of the loss as a percentage of EAD in the event of a loan default.

  Whilst there has been considerable focus on PD and LGD modelling in the banking industry and academia over the last decade, EAD modelling remains a somewhat neglected field of research, despite there being a range of commercial applications for models that accurately predict exposure in the event of a default occurring.

  The aim of this project is to investigate and assess alternative statistical modelling techniques applied to the problem of predicting EAD on retail and corporate lending portfolios.

- **Acquisition Profitability Optimization**

  This topic relates to optimizing processes used to make decision on new loan applications by using various statistical modelling techniques. Most banks make decisions on whether to accept or reject an application for a new loan

based on automated, model-driven processes. These processes rely on accurate credit scoring models, which predict default propensity, and also a second set of models commonly referred to as "cut-off models" designed to select the minimum credit score criteria (aka the "cut-off") which optimizes portfolio profitability. This project will focus on enhancing the cut-off models by reviewing a range of different approaches and modelling techniques.

- **Credit Conversion factor**

  A credit conversion factor (CCF) is an estimate of how much of the undrawn amount of an account the borrower will use over the 12 months prior to default. Currently in Australia for wholesale borrowers a CCF of one is required when calculating regulatory capital, evidence suggests this is very conservative. Initial analysis shows the credit conversion factor is a function of the account type, time to default, borrower characteristics and the proportion of the account undrawn 12 months prior to default. There is an opportunity to model this behaviour using Westpac data.

- **State of the cycle model**

  When calculating capital an estimate of the Probability of Default (PD) is required. The PD should be a through the cycle estimate and include a downturn. One key problem is that Australia has not experienced a downturn in the last 20 years. In order to understand how the PD will vary during a downturn a state of the cycle model is required. The intention of the state of the cycle model is to link the wholesale default rates to economic indicators using international and domestic economic data. Creating a model will allow forecasting and back casting to be performed and a through the cycle PD to be calculated.

How to apply

Please see the instructions on this page: www.maths.unsw.edu.au/currentstudents/honours-mathematics-and-statistics