PhD opportunities in Statistics at St Andrews, 2020-2021

(Updated 6th October 2019.)

Applications are welcomed for students wishing to undertake a PhD in Statistics at St Andrews. Full funding (fees, plus stipend of approx. £15,000) is available for well-qualified students; we encourage applications as soon as possible to maximize your chances of being funded. UK, EU and other overseas students are all encouraged to apply. New PhD students would typically start in September 2020, but this is flexible.

Some general information about the division of statistics is given below, followed by a list of specific topics that are on offer this year. Finally, more information is given about how to apply.

Statistics at St Andrews

Statistics is a lively area of research at St Andrews. The Division of Statistics is one of three within the School of Mathematics and Statistics (https://www.mcs.st-and.ac.uk), and consists of 13 members of academic staff, 12 research staff and 11 PhD students. It was ranked first in Scotland in the 2008 Research Assessment Exercise (the last for which statistics was ranked separately from mathematics); the School of Mathematics and Statistics was ranked first in Scotland in the 2014 Research Assessment Framework, and is ranked 3rd in the UK for 2020 by the Complete University Guide (source https://www.thecompleteuniversityguide.co.uk).

One major research strength is in the area of statistical ecology: contained within the School is the world-leading Centre for Research into Ecological and Environmental Modelling (CREEM; https://www.creem.st-and.ac.uk), which is housed in tailor-made facilities at the St Andrews Observatory on the edge of the town. We are a founding member of the National Centre for Statistical Ecology (https://www.ncse.org.uk), a multi-institution consortium that ensures regular intellectual exchange between researchers worldwide with similar interests. Several members of CREEM are also part of the university’s multi-school Centre for Biological Diversity (CBD; https://synergy.st-andrews.ac.uk/cbd/).

A second more recent and rapidly developing research focus is statistical medicine and molecular biology group (https://sites.google.com/view/smmb/home), led by a joint professorial appointment (Prof. Andy Lynch) between the Schools of Mathematics and Statistics and the School of Medicine, a recent appointment at the interface between statistics and mathematical biology (Dr. Giorgos Minas), and research interests from several other staff members.

Many staff members are also active more generally in the field of statistical inference. Research areas include Bayesian statistical inference and, relatedly, computer-intensive inference, data mining, data smoothing, latent state models and experimental design.

A brief summary of the research interest of each member of staff is given at the bottom of this section; more details can be found on the school and CREEM web sites.
New PhD students join a high-calibre but friendly research environment. Training is provided in the first year in as part of St Andrews’ participation in the Scottish Mathematical Sciences Training Centre (https://smstc.ac.uk/) and Academy for Postgraduate Training in Statistics (www.apt.ac.uk), the latter consisting of four one-week residential courses. Students may get the opportunity to become involved in externally-funded research as part of CREEM’s consultancy group (https://www.creem.st-andrews.ac.uk/consultancy/); they may also be able to assist on statistics training workshops delivered to professional scientists both in the UK and abroad. Some PhDs are supervised jointly with scientists from other institutions, and there may be opportunities for study at those places. PhD studies are expected to last approximately 3.5 years.

St Andrews is a small, vibrant university town. It is situated on the east coast of Scotland and framed by countryside, beaches and cliffs. The town has a rich cultural heritage, having once been at the centre of Scotland’s political and religious life. Today it is known around the world as the Home of Golf and a bustling student town with a distinctively cosmopolitan feel, where students and university staff account for more than 30% of the local population. The university is the oldest in Scotland and third oldest in the English-speaking world. It is the top-rated university in Scotland for teaching quality and student satisfaction, and among the top rated in the UK for overall research; it regularly comes in the top few places in UK league tables compiled, for example, by broadsheet newspapers (e.g., 1st place 2020 Times and Sunday Times; 2nd place 2020 The Guardian). Its international reputation for delivering high quality teaching and research and student satisfaction make it one of the most sought-after destinations for prospective students from the UK, Europe and overseas.


**Brief summary of academic staff interests in the Division of Statistics**

- Rosemary Bailey – design of experiments in agriculture, horticulture, ecology and medicine
- David Borchers – spatial capture-recapture, wildlife surveys, spatial modelling
- Carl Donovan – data mining, commercial statistics, multivariate statistics
- Steve Drasco – gravitational waves, black holes, machine learning
- Andy Lynch – design or analysis of molecular biology experiments, especially applications of DNA/RNA sequencing to cancer research
- Giorgos Minas – statistics in molecular biology and medicine
- Michail Papathomas – Bayesian methods with application to genetics and biostatistics
- Valentin Popov – time series and hidden process models
- Chris Sutherland (from June 2020) – statistical ecology
- Len Thomas – wildlife (particularly acoustic) surveys, population dynamics modelling
- Hannah Worthington – Hidden Markov models for mark-recapture analyses, links between ecological and epidemiological/medical methods

Academic staff not taking PhD students in the coming academic year:
Specific projects offered for 2020-21

We are currently looking for candidates for the following projects. In addition, prospective candidates with general interests related to those of staff members (see above) are welcome to contact them to discuss other possible projects.

Design of Experiments.
Supervisor: Rosemary Bailey

A topic in design of experiments, to be decided between the student and the supervisor. It could focus on methods for constructing particular types of design; methods of, and problems in, randomization; two-phase experiments; experiments in a particular area of application.

Statistical models for digital ecological surveys
Supervisor: David Borchers

Ecological surveys are the foundation on which evidence-based conservation of the planet’s biodiversity and wildlife resources is built. Wildlife surveys that provide the evidence base have traditionally been conducted by humans and the resulting data treated as “snapshots” in time. However, digital survey devices like camera traps, acoustic arrays or aerial video platforms generate streams of data, not snapshots. These data are more appropriately viewed as time-to-event data, with the events being detections of the species of interest. Using the event times has the potential to yield much richer inferences about the populations under study than snapshot data can, but very little work has been done in this area. If you are interested in developing and applying new statistical methods to exploit the power of digital survey devices, this may be the PhD for you. Specific applications include camera trap surveys of large cats, acoustic surveys of gibbons, chimpanzees, and various bird species, drone-borne aerial surveys of snow leopard prey, and many others.

Modelling encounters in surveys of unmarked animal populations
Supervisors: David Borchers, Richard Glennie (University of St Andrews) and Marcus Rowcliffe (Institute of Zoology, Zoological Society of London)

Many wild animal populations are monitored by placing detectors (e.g. cameras or microphones) in a study area and recording encounters with individual animals. From these encounters, the goal is to estimate population density. A common problem is that individuals are not identifiable, often termed “unmarked”, so we cannot know which individuals were seen in each encounter.
Random encounter models are a method to estimate population density from unmarked populations by using auxiliary information on how individuals move: if you know how individuals move, you can estimate how many times a single individual would be encountered, and so deduce how many individuals produced the total number of encounters observed.

This PhD project will focus on developing random encounter models (REM) in one or more of the following ways: (1) construct a maximum likelihood based framework for estimation; (2) incorporate alternative models for animal movement; (3) extend density and encounter models to vary, with correlation, spatially and temporally; (4) build joint models for partially marked populations.

Key references:

*Spatial capture-recapture methods for snow leopards.*

Supervisors: David Borchers, Richard Glennie (University of St Andrews) and Koustubh Sharma (Snow Leopard Trust)

We do not know how many snow leopards are left in the world. Snow leopard range country governments and scientists have launched an ambitious initiative to develop a robust assessment of the global snow leopard population. Spatial capture-recapture (SCR) methods are central to these efforts. The very heterogeneous nature and massive range of suitable snow leopard habitat, the tiny fraction of the range that can be surveyed in any year, and the variety of data types (camera trap data, GPS tag data, genetic sampling data, prey survey data, environmental data) that are available for informing estimates of abundance and density, raises methodological challenges for statistical analysis and for survey design. In collaboration with the Snow Leopard Trust (SLT), and using data provided by the SLT, this PhD will develop statistical methods to address some of these challenges. This might involve developing open- or closed-population methods that integrate a variety of data sources, developing models that use times of detection to draw inferences about activity patterns and habitat use, incorporating uncertainty in identifying individuals from photographs, and integrating automated identification methods into inference.

*Object classification from mobile and static sensor feeds*

Supervisor: Carl Donovan
The demand for video processing is rapidly increasing, driven by greater numbers of sensors with greater resolution, new types of sensors, new collection methods and an ever wider range of applications. For example, video surveillance, vehicle automation or wildlife monitoring, with data gathered in visual/infra-red spectra or SONAR, from multiple sensors being fixed or vehicle/drone-mounted.

This project will focus on a specific application – object (animal) extraction and classification from extremely high-resolution aerial video from moving platforms. Issues of data size, dynamic backgrounds, rapid platform and target movement and classification errors will all need to be resolved and propagated into the final goal – inferring the densities of target species.

The project will require solving substantive computational bottlenecks and creative programming e.g. GPU and distributed file systems. Elements can be found in Erichson & Donovan (2016), but is only a tiny fraction of what is required.

References:

Trading in Peer-to-peer (P2P) markets
Supervisor: Carl Donovan

Peer-to-peer market trading is becoming more prevalent any many areas that were traditionally the domain of large companies – stocks, foreign exchange, gambling. This project will look at a variety of statistical fundamentals for these areas including the practical application of statistical arbitrage, traditional arbitrage, algorithmic and high-frequency trading.

The project deals with large data issues, data-mining/machine-learning methods, cloud-computing and the dynamics of automated trading via APIs. The project will be heavily computational in either R or python, with potentially compiled languages for computational bottlenecks. The ideal candidate would have a good grasp of practical computing and statistics.

Automated evaluation of geo-political risk
Supervisor: Carl Donovan

There are massive amounts of data presented to the internet in real or near-to-real-time that allow monitoring of economic and societal conditions, amongst other things. This has been used to some effect in the automated monitoring of stock fluctuations, used to inform algorithmic trading. The evaluation of data-sources and curation of these on the basis of predictive power is an area requiring exploration. The project here will focus on a multitude of data sources that would allow the real-time evaluation of geo-political conditions around the globe, with the intention to predict various market shifts and impending political flash-points. Data will be captured by a wide range of sources, including multiple languages, print, audio and video. Text-mining methods will be developed to generate topic models and monitoring these over time, including those emergent. Methods will be
explored that evaluate predictive performance for the purposes of curating data-sources and selection of modelling techniques.

The project deals with large data issues, data-mining/machine-learning methods, cloud-computing and interaction with APIs. The project will be heavily computational in either R or python, with potentially compiled languages for computational bottlenecks. The ideal candidate would have a good grasp of practical computing and statistics.

**Investigating modes of action of genetic risk variants through integrated analysis of multiple high-dimensional “omics” data.**

Supervisors: Andy Lynch and Michail Papathomas

Genome-wide association studies have identified thousands of genomic loci that are associated with higher risk of a trait (often a disease such as breast cancer). While these associations may have been identified, the mechanisms through which they act have tended not to be elucidated. This despite the growing number of diverse data sets potentially available for the purpose (see for example The American Journal of Human Genetics 103, 637–653 for discussion).

This project will look to develop a flexible modelling framework to incorporate many potential sources of evidence in suggesting and evaluating mechanisms of action. This could potentially build upon work of Papathomas et al. (2012) and related efforts in producing a flexible Bayesian approach for the analysis of GWAS data. The evaluation of potential mechanisms may then feed back into the detection and prioritization of association loci through, e.g., specification of prior probabilities.

Reference:

**Adapting methods from statistical ecology and applying them to molecular data in cancer**

Supervisors: Andy Lynch and Hannah Worthington

There are a number of questions that arise from proteomic and RNA studies in cancer that have clear parallels in the world of ecology. This is particularly the case as more frequently tumours are either sampled multiple times spatially (in either a designed or convenience fashion) or are analysed in a non-destructive manner that preserves all spatial data (but sometimes for a greatly reduced number of molecule types).

These questions include

- How many molecular 'species' are present?
- What is the abundance of a particular species in a sample for which we have imperfect sampling, and what can we infer about the dynamics?
• To what degree do two species have the opportunity to interact?
• etc.

This project will look to develop and apply methods to answer these and similar questions, drawing on the large corpus of work addressing analogous problems in statistical ecology.

**Bayesian identifiability for log-linear models.**

Supervisor: Michail Papathomas

Log-linear modelling is the standard approach for investigating the full joint dependence structure between categorical variables such as phenotypes and SNPs. Complex dependence structures can be easily discerned using graphical log-linear models (Papathomas and Richardson, 2016). This can potentially lead to identifying functionally important pathways. The number of cells in the associated contingency table increases rapidly with the number of variables, creating sparse contingency tables with a number of zero cell counts, even for a large number of subjects. The presence of zero cell counts can potentially make some model parameters non-estimable, also referred to as non-identifiable. Non-identifiability is a major impediment to evaluating how risk factors interact, and understanding important biological mechanisms. Problems associated with identifiability are currently not sufficiently understood, and have not been addressed in a systematic manner. The aim of this project is to develop methods that will utilize information pertaining to the Bayesian identifiability of interaction parameters, towards choosing the best log-linear model given the data.

Reference:

**Determining the origins of galaxy bimodality using hierarchical Bayes methods**

Supervisors: Vivienne Wild (Physics), Michail Papathomas

This project is funded by the UK research council STFC. To be eligible for STFC funding, we strongly encourage applications before February 1st, with interviews being held in late February and early March.

How galaxies form and evolve is one of the outstanding questions of modern astrophysics. We have known for many decades that massive galaxies come in two main types - elliptical/quiescent and spiral/star-forming. However, it remains largely unknown why some galaxies are still forming stars while others are "red and dead". Extremely large galaxy surveys are providing an increasingly detailed census of both local and distant galaxies. Considerable progress is being made on quantifying the changing demographics of the galaxy population over the majority of the age of the Universe, but significant improvements in methods are required to dramatically improve our understanding of the physics behind the observable properties of galaxies.
In the last decade a Bayesian approach to the fitting of sophisticated models to high quality spectra and/or multiwavelength photometry has become common place in the analysis of galaxy spectral energy distributions (SEDs) at all redshifts (Walcher et al. 2011). The result is robust physical properties, such as galaxy stellar mass, dust content and star formation history, together with well quantified degeneracies between these parameters. However, by treating galaxies as independent entities to determine their physical properties, we are missing vital population information. A better approach would be to treat galaxies as a population of objects with a common origin and common underlying variables. This could tighten constraints on the physical properties of individual galaxies as well as the underlying relationships that impact the life of galaxies. It could also allow us to extract information from larger surveys with lower quality observations.

Hierarchical Bayes techniques have been used in the astronomical literature to solve problems as diverse as quasar redshift estimation (Bovy et al. 2011), exo-planet orbit analysis (Hogg et al. 2011), properties of supernovae light curves (Mandel et al. 2009), and photometric redshifts (Leistedt et al. 2016). They differ from standard Bayesian methods by fitting the entire dataset in a coherent manner, instead of single objects as entirely independent entities. By applying these methods to galaxy evolution studies, we will improve our ability to break degeneracies between physical parameters and understand the underlying processes governing galaxy evolution. These methods could be applied to e.g. complete populations of galaxies in spectroscopic or photometric surveys, or entire integral field datacubes of single galaxies.

This interdisciplinary project will be jointly supervised by Drs Vivienne Wild and Michail Papathomas in the Schools of Physics and Astronomy and Mathematics and Statistics respectively. Dr Wild has built her career around developing and applying novel statistical techniques to explore datasets on galaxy evolution, focussing most recently on understanding the nature of post-starburst galaxies. Dr Papathomas is an expert in Bayesian modelling, both in the development of new methods and their application to a wide variety of datasets.

Large extragalactic datasets are already available for analysis, both at low and high redshift. The School of Physics and Astronomy is a member of the UK participation group in SDSS-IV, the fourth generation of Sloan Digital Sky Surveys, a large international collaboration encompassing several astronomical surveys. The methods developed during the project could also be applied to upcoming datasets to which the group has proprietary access such as DESI bright galaxy survey, LSST science verification data and VLT/MOONS near-infrared spectroscopic survey.

The project involves the development of statistical techniques to make them applicable to astronomical datasets. This project would suit students with a background in (astro)physics but aptitude for maths and/or statistics, and students with a background in maths or statistics and interest in astrophysics.

For more information please contact Dr Vivienne Wild and Dr Michail Papathomas (vw8@st-andrews.ac.uk, M.Papathomas@st-andrews.ac.uk).

References:
Modelling local population dynamics for whale sharks in the Maldives

Supervisor: Hannah Worthington

Data for this project will be provided by the Maldives Whale Shark Research Programme. The broad aim is to further the understanding of the local population dynamics of this enigmatic species.

Potential areas for investigation include:

- Robust stopover models for long-term mark-recapture data. The extension of these models to include biologically realistic behaviour (temporary migration, heterogeneous captures, individual covariates, environmental covariates). Robust estimation of recruitment probabilities and stopover duration on different temporal scales.

- Daily estimation of abundance and/or animal density. Modelling seasonal variation caused by, for example, lunar cycles, monsoon season, time-varying environmental conditions.

- Modelling animal movement and identifying habitual behaviour patterns. Investigating differences in behaviour for sharks of different ages or regions.

- The integration of different sampling methods including emerging technologies (drone surveys, sonar, eDNA). The inclusion of citizen science data.

- The potential extension of existing mark-recapture methods to a continuous-time framework.

Application procedure

Although there is no fixed deadline (unless noted otherwise for a particular topic), you are strongly encouraged to make your application as early as possible!

Many details of the general requirements and admissions procedure are given at the university website [https://www.st-andrews.ac.uk/study/pg/apply/research/](https://www.st-andrews.ac.uk/study/pg/apply/research/)

Applicants should have a good first degree (UK class 2:1 or better, or international equivalent) in mathematics, statistics or another scientific discipline with a substantial numerical component. Applicants with degrees in other subjects, such as biology should have the equivalent of A-
level/Higher mathematics, and experience using statistical methods. Such applicants are invited to
discuss their qualifications with the Postgraduate Officer (contact details below). A masters’ level
degree (MSc, etc.) is an advantage, as is any other relevant professional experience. Please note
that our primary criterion for selection is academic excellence; most successful applicants
(particularly those who are awarded scholarships) have a good to very good 1st class undergraduate
degree and/or a distinction at MSc level. Those who do not have English as a first language, and who
have not undertaken an undergraduate or graduate degree taught in English, should provide
evidence of English proficiency (minimum IELTS 6.5 or equivalent).

Potential applicants are encouraged to contact the Postgraduate Officer responsible for PhDs in
Statistics, in advance of making a formal application. He is: Len Thomas, email
len.thomas@st-andrews.ac.uk, tel. 01334 461801.

To make a formal application, complete the appropriate online form at https://www.st-
andrews.ac.uk/study/pg/apply/research/ (click on “Apply Now” on that page). You also need to
provide the following supporting documentation: CV, evidence of qualifications and evidence of
English language (if applicable). You are welcome to include a covering letter. You don’t need to
provide a research proposal unless you are proposing your own project, or sample of academic
written work. You will need to ask two referees to provide academic references for you – once you
fill in their name on the form, they will be sent emails asking them to upload their references.
Please note that we give serious consideration to both the stature of your referees and the remarks
that they make about you. More details about the application procedure are given at
https://www.st-andrews.ac.uk/study/pg/apply/research/

Further School-specific information is on this page
https://www.st-andrews.ac.uk/maths/prospective/pg/phdprogrammes/
and in this pdf
https://www.st-andrews.ac.uk/media/school-of-mathematics-and-
statistics/documents/prospective-students/st-andrews-mathsstats-pgr-info.pdf
which also contains some information about funding and scholarships. In addition to the
scholarships mentioned there:

- The Centre of Research into Ecological and Environmental Modelling has a small scholarship
  fund; all students applying for School funding with an intended PhD topic in the field of
  statistical ecology are automatically considered.
- An up-to-date list of external scholarships is given at https://www.st-
  andrews.ac.uk/study/fees-and-funding/postgraduate/scholarships/research-scholarships/.

We look forward to hearing from you!