Adaptive control of stochastic queueing networks

The evolution of queuing systems is often random, and key system variables/parameters may be unknown or only partially observed. Providing a stochastic model for these systems with the goal of improving efficiency or forecasting and bringing them under online control, leads to reducing the customer waiting times, better server utilization, and stability. The main goal of this project is to devise an appropriate and optimal model for a network of queues that fits practical applications in the fields of biology, health services, energy, manufacturing, traffic and communication networks.

Requirements: Very good grade in STATS 225 and 320 or equivalent. Programming skills would be an advantage.

Contact: Azam Asanjarani (azam.asanjarani@auckland.ac.nz)

Stochastic modelling of patient's flow in a hospital

Since the outbreak of a new epidemic in 2020, the importance of managing patient flow within a hospital has become more widely recognised. Almost every country is dealing with a severe shortage of healthcare supply as a result of unexpected and unregulated patient inflow into hospitals, which has not only resulted in a significant decline in overall healthcare system performance but has also put patients' safety at risk. The goal of this project is to build a stochastic model for predicting an individual's progression through various stages of a disease using Markov decision processes and simulation methods. Also, make a reasonable contribution to the development of a prediction model that predicts the risks and chances of patients' expected trajectories through different departments of a hospital.

Requirements: the student needs to be familiar with stochastic processes such as Markov chains and queueing systems and have good programming skills.

Contact: Azam Asanjarani (azam.asanjarani@auckland.ac.nz)

Moment matching problem for truncated multivariate distributions

The matching of distributional parameters to obtain desired moments is an intriguing classic problem in statistics and econometrics. The application of truncated distributions occurs frequently in a wide range of scientific problems. The goal of this project is to solve the moment matching problem with a novel dynamic method designed specifically for truncated multivariate distributions.

Requirements: Mathematics skills (e.g. proofs, limits) is essential. Some knowledge of basic probability and stochastic processes is recommended (STATS 125, STATS 325, STATS 320).

Contact: Azam Asanjarani (azam.asanjarani@auckland.ac.nz)

Scheduling for a processor sharing system

In a variety of real-life queueing systems such as manufacturing, telecommunication, transportation, supermarkets or hospitals, job requests arrive continuously and the servers (e.g. machines, cashiers, doctors …) may not immediately supply their customers with the amount or type of service they required. In these cases, we use scheduling policies to determine which requests in the queue are serviced at any given time, how much time is spent on each, and what happens when a new request arrives. The result would be reducing the waiting time in the queue and treating each request fairly. The aim of this project is to solve the problem of scheduling arrivals to a congestion system (such as a traffic intersection) with a finite number of users and identical deterministic demand sizes.

Requirements: Very good grade in STATS 225 and 320 or equivalent.

Contact: Azam Asanjarani (azam.asanjarani@auckland.ac.nz)

Joint autoregressive modelling of stream flows and barometric pressure in the South Island.

This BSc(Hons) or 45pt MSc project will use two main datasets: water inflows to South Island lakes and rivers and observations of air pressure at South Island meteorological sites. While the inflows are of more immediate practical interest -- they determine the availability of water for hydropower generation -- the air pressure is expected to be less variable, better behaved, and generally easier to model.

The aim of the project is to construct joint time series models of these phenomena. Allowing for seasonal variation will be a significant complication, with the inflows in particular exhibiting seasonal variation not only in mean, but also in variance and autocorrelation structure. Multivariate models are the eventual goal, to combine the two main variables and to model multiple sites.
Requirements: Concurrent or previous enrolment in STATS 326/726/727 would be an advantage, but is not a requirement.

Contact: Geoffrey Pritchard. (g.pritchard@auckland.ac.nz)

Statistics Education: Markov processes – a visual approach for learners

Researchers in the field of probability education research are calling to reform the teaching of probability from a traditional mathematical base to include more emphasis on modelling using technology. Lecturers in STATS 125 have been using a prototype tool when teaching Markov processes. This project will investigate what new understandings students may gain from learning Markov processes through the use of technology.

An important component of this project will involve conducting a literature review. This means that skills in reading and critiquing research papers and essay writing are important.

Requirements: Good grades in STATS 125 and STATS 210. It is also desirable that you have marking and/or tutoring experience in at least one of STATS 125 or STATS 210.

This project would suit someone with a genuine interest in Statistics education.

Contact: Heti Afimeimounga (h.afimeimounga@auckland.ac.nz), Azam Asanjarani (azam.asanjarani@auckland.ac.nz) and Stephanie Budgett (s.budgett@auckland.ac.nz)

Genealogies of samples from stochastic populations and biodiversity models

This project in probability theory will investigate some stochastic models for biodiversity and genealogical trees for samples of individuals chosen at random in stochastic population models. In particular, some neutral models for extinctions and speciations using branching processes will be investigated mathematically, including the genealogical structure of reconstructed phylogenetic trees. For example, see Gernhard (2008), Stadler (2009), and Harris, Johnston, Roberts(2020). The project will include directed reading for any necessary background material in probability, such as Markov chains, branching processes, and Poisson processes. Computer simulations can optionally be used to exhibit typical behaviours and theoretical results.

Requirements: A good background in probability (eg. Stats125, Stats225) and very good mathematics (eg. proofs, limits, calculus, differential equations) is essential. Some more advanced knowledge of stochastic processes or Markov chains is also strongly recommended (eg. Stats325, Stats320).

Contact: Simon Harris (simon.harris@auckland.ac.nz)

Inhomogeneous branching Brownian motions

Brownian motion is a fundamental model in modern probability theory for the random diffusion of a particle, and can be thought of as the natural scaling limit of the well known probabilist's simple random walk. Branching Brownian motions are population models in which each particle currently alive independently moves around in space as a diffusion, but also gives birth to offspring at random during its lifetime. This project will investigate some inhomogeneous branching Brownian motions, where the motion, branching rates and death rates depend on current spatial position (or time) of the particles. Some fundamental questions include survival probabilities and how quickly the population colonises space given it survives. Probabilistic results about Branching Brownian motions can also yield results in mathematical analysis about corresponding reaction-diffusion equations (non-linear partial differential equations). For example, see Harris & Harris (2008), Berestycki, Brunet, Harris et al. (2010, 2017).

Requirements: A good background in probability (eg. Stats125, Stats225) and strong mathematics (eg. proofs, limits, calculus, differential equations) is essential. Some more advanced knowledge of stochastic processes or Markov chains is also strongly recommended (eg. Stats325, Stats320).

Contact: Simon Harris (simon.harris@auckland.ac.nz)

Title: Saliency Maps for Data Visualisation

A visual saliency map predicts where attention will be directed when viewing an image. This is relevant to data visualisation because we may want to direct attention to specific elements of a plot; we may also want to avoid diverting attention to unimportant elements within a plot.

This project aims to develop an R package to generate visual saliency maps for R plots.

The project will involve background reading of the literature of visual saliency maps, R code design and development to generate saliency maps, and some data wrangling to work with data sets containing test images and the results of eye-tracking experiments.

Requirements:
This project requires a student to be very strong in R programming and data visualisation. It would suit someone who likes to write code in their own time.
Good marks in STATS 220 and/or 380 are essential.
Good marks in STATS 782 and/or 787 would be ideal.
A background in psychophysics would be a bonus.

Contact: Paul Murrell (p.murrell@stat.auckland.ac.nz)

Expanding the 'gridGeometry' package

The 'gridGeometry' package provides an interface to the 'polyclip' package. 'gridGeometry' makes it easy to draw complex shapes in 'grid' by combining simple shapes (using operators like union, intersection, and set minus). One possible application is the creation of customised data symbols for use in a scatterplot.

There are features within the 'polyclip' package, such as generating offset polygons, for which 'gridGeometry' does not currently provide an interface.

The aim of this project is to expand the 'gridGeometry' package to provide an interface to more of the features within 'polyclip'.

The project will involve learning about how the 'gridGeometry' package works, followed by R code design and development to expand the 'gridGeometry' package.

References: https://stattech.wordpress.fos.auckland.ac.nz/2019/03/04/2019-01-a-geometry-engine-interface-for-grid/

Requirements: This project requires a student to be very strong in R programming.
It would suit someone who likes to write code in their own time.
Good marks in STATS 220 and/or 380 are essential.
Good marks in STATS 782 and/or 787 would be ideal.
Familiarity with Linux and git would be advantageous.

Contact: Paul Murrell (p.murrell@stat.auckland.ac.nz)

Bayesian inference for failure times of load-sharing systems with damage accumulation

In many engineering applications, it is of interest to test the reliability of systems which are composed of parallel components. In a load-sharing system, the stress is redistributed to the surviving components after one of the components fails. This project aims to develop a Bayesian alternative to the existing maximum-likelihood approach for modeling the intensity function of a point process and apply this to real failure time data of pre-stressed concrete beams that are each made up of several tension wires.

Requirements: Good knowledge of applied Bayesian methods e.g. from STATS331 or STATS 731, good programming skills and knowledge of R, JAGS/WinBUGS are essential.

Contact: Renate Meyer (renate.meyer@auckland.ac.nz)

The inflated density ratio approach for Bayesian model selection.

One of the most popular techniques used for model selection is the Bayes factor. This is defined as the ratio of the marginal likelihoods of the models that are being compared. The marginal likelihood is the normalising constant in the Bayes theorem and in general must be approximated through numerical methods. Techniques such as thermodynamic integration, stepping-stone sampling, and nested sampling are used to estimate it in Bayesian Gravitational Waves (GW) data analyses. However, these techniques require a high computational cost. In this project, we aim to study the inflated density ratio approach. This method is less computationally expensive and can potentially be the simplest and most accurate method to estimate the marginal likelihood for GW models.

Requirements: Good grade in STATS331 or STATS731, good computing skills, (no experience with gravitational wave data analysis required).

Contact: Patricio Maturana Russel (p.russel@auckland.ac.nz) or Renate Meyer (renate.meyer@auckland.ac.nz)

Bayesian estimation of the long-term linear trend of New Zealand annual average temperatures

An accurate understanding of the long-term evolution of the temperature is key to understanding the impact of global warming. NIWA provides time series of annual average temperatures at various different sites in New Zealand. However, these temperature time series lack homogeneity due to changes in instrumentation and re-siting of recording stations that has necessitated adjustments in the past. The goal of this project is to use a hierarchical Bayesian model to estimate the slope. It will aim for a robust analysis by using a nonparametric approach to model the time series errors. The first phase of the project will be concerned with data wrangling, accessing the data from the NIWA website and bringing it into a suitable format for subsequent analysis using R packages for nonparametric time series errors. Exploratory analysis will be part of this phase. In a second phase, the model for time series errors will need to be combined with a hierarchical linear model for the slope. Sampling from the posterior distribution will either be performed using JAGS and/or Metropolis-Hastings routines written in R.

Requirements: A good knowledge of and interest in Bayesian inference, MCMC techniques, and time series as well as good programming skills and knowledge of R and JAGS are essential.

Contact: Matt Edwards (medw076@aucklanduni.ac.nz),Renate Meyer (renate.meyer@auckland.ac.nz)

Visualising the saddlepoint approximation

The saddlepoint approximation is a systematic method for approximating an unknown density function in terms of a known moment generating function. It is useful when each individual in a large population contributes to a single random variable, and has often been used in statistical ecology.

While the method has its roots in approximations for integrals, the saddlepoint approximation can also be given a statistical interpretation. Indeed, the saddlepoint approximation is a kind of normal approximation, applied after a key "tilting" step. In this project, you will develop an R Shiny tool to visualise the tilting step of the saddlepoint approximation, so that users can better assess when the saddlepoint approximation is likely to give good results.

Requirements: This project will primarily need a familiarity with R Shiny programming - the mathematical aspects of the saddlepoint approximation are not prerequisites. A sense for how to present distributional information visually would be a plus.

Contact: Jesse Goodman (jesse.goodman@auckland.ac.nz)

Improving the accuracy of the saddlepoint approximation for count data

The saddlepoint approximation is a systematic method for approximating an unknown density function in terms of a known moment generating function. It is useful when each individual in a large population contributes to a single random variable, and has often been used in statistical ecology.

The saddlepoint approximation works best for densities, when the underlying random variable is continuous. For discrete random variables, the traditional saddlepoint approximation works less well, and always fails at the boundary. This project will implement several new alternative saddlepoint approximations in a simple model and assess how these proposed alternatives compare to existing methods.

Requirements: Experience with R programming and simulation would be a plus. The mathematical aspects of the saddlepoint approximation are not prerequisites, but mathematical applications could be explored as part of the project depending on the student.

Contact: Jesse Goodman (jesse.goodman@auckland.ac.nz)

Huber loss for forecasting hierarchical time series

Multivariate time series with linear constraints is known as hierarchical time series. Forecasts from these structures need to satisfy the same linear constraints as the data. Addressing this problem, the methods proposed in the literature follow a two-step process. First, it forecasts each series in the structure independently using any appropriate univariate time series model. Second, it maps these forecasts using projection matrices. The theory developed for these methods is based on L2-norm. This study investigates the impact of using Huber's loss function for forecast reconciliation.

Requirements: Good programming skills in R, and good understanding of linear algebra. Prior knowledge on time series analysis is an advantage.

Contact: Shanika Wickramasuriya (s.wickramasuriya@auckland.ac.nz)

NOVELIST covariance estimator for forecasting hierarchical time series

Forecasting a collection of time series with aggregation constraints has attracted much attention among forecasting practitioners. Among the methods available in the literature, MinT is commonly used. However, its forecasting performance is greatly influenced by the estimate of the covariance matrix used. A shrinkage estimator with a diagonal target is widely used in practice. This study intends to assess the performance of the NOVELIST covariance estimator proposed by Huang and Fryzlewicz (2019) on forecast reconciliation.

Requirements: Good programming skills in R, and good understanding of linear algebra. Prior knowledge on time series analysis is an advantage.

References: Huang, N. and Fryzlewicz, P. (2019). NOVELIST estimator of large correlation and covariance matrices and their inverse. Test 28:694-727.

Contact: Shanika Wickramasuriya (s.wickramasuriya@auckland.ac.nz)

AI powered expense report system

Corporates use “Expense Report” to code or reimburse a group of spends/transactions. This is what we call Expense Report, and it’s very popular in the US region.
Patterns can be learned from past behaviour, such as, by dates that transactions happened, or Merchant, or locations, etc.
In this project we are working on machine learning algorithms to learn these kinds of patterns and provide recommendations to the users when they use “Expense Report systems”. This project will be a collaboration between industry and University.

Requirements: Strong computational and algorithmic skills are essential for this project.

Contact: Mehdi Soleymani (m.soleymani@auckland.ac.nz)

Generally-Altered, -Inflated, -Truncated and -Deflated Regression, With Application to Heaped and Seeped Counts

Zero-altered, -inflated and -truncated count regression are now well established, especially for Poisson and binomial parents. Recently these methods were extended to Generally-Altered, -Inflated, -Truncated and -Deflated Regression (GAITD regression) and implemented in the VGAM R package for three 1-parameter families and one 2-parameter family. In GAITD regression the four operators apply to general sets rather than {0}. Also,the four operators may appear simultaneously in a single model.
Elements of the four mutually disjoint sets of support values are called 'special'. Parametric and nonparametric variants are proposed: the latter based on the multinomial logit model (MLM), and the former on a finite mixture of the parent distribution on nested or partitioned support. The resultant "GAITD Mix-MLM combo" model has seven special value types. GAITD regression offers much potential for the analysis of heaped (digit preference due to self-reporting) and seeped data.

This project is consolidate the above and to investigate some extensions. Some specific examples include:

1. Find new data sets from a wide range of fields exhibiting heaping and seeping. Perform some analyses.

2. Find any bugs in the software. Suggest any improvements (such as initial values) and additions.

3. Marginal effects: extend margeff() to compute the first derivatives of the MLM terms.

4. Find data sets that are underdispersed with respect to the Poisson. Apply the GT-Expansion method of analysis.

Requirements: Ideally, a student working on this project would have strong computational skills and a solid understanding of generalized linear models (GLMs; e.g., STATS 330 & 310).

Contact: Thomas Yee (**t.yee@auckland.ac.nz**)

VGAMviz

This project in statistical computing entails writing an R package called VGAMviz for improved VGAM plots. The VGAM objects are fitted in package VGAM, and the idea is to mimic what mgcViz does to mgcv.
Some exposure to regression modelling involving smoothing would be good background. This project will use S4 object-oriented programming and graphical software such as the ggplot2 package to design plotsof regression objects such as those involving smoothing splines. There are other ideas worth exploring, such as dynamic/interactive graphics and big data sets.

Prerequisites: STATS 330, and STATS 380 or 782

Contact: Thomas Yee (**t.yee@auckland.ac.nz**)

Analysis of Fly Fishing Data

This project is ideally needs somebody familiar with (freshwater) fly fishing and has good R programming skills. Much data processing is first required to get the data into shape, and it is essential to know about the main fly fishing techniques such as nymphing, wetlining, dry fly, etc. The data was collected by the Department of Conservation over a long period of time, and this work is in collaboration with a DOC scientist.

Requirements: knowledge about fly fishing, STATS 330 and good R programming skills.

Contact: Thomas Yee (**t.yee@auckland.ac.nz**)

Multinomial Logit Model

Abstract: The aim of this project is to improve the multinomial() family function in the VGAM R package. The multinomial logitmodel is the standard model for regressing a nominal categorical response against a set of explanatory variables. It can suffer from numerical problems with sparse data, however, bias reduction can be a solution for this (Ding and Gentleman, JCGS, 2005). One task is to implement this within the function. Also, we could writefunctions to conduct a score test, as well as the Hausman-McFadden test for independence of irrelevant alternatives (IIA). Time permitting, another useful feature would be to handle the nested multinomial logit model, however this would be quite a challenge.

Requirements: This project would suit a student with good R programming skills and has done STATS 310 and STATS 330.

Contact: Thomas Yee (**t.yee@auckland.ac.nz**)

Multivariate count distributions estimated by iteratively reweighted Poisson regressions.

Zhang et al. (2017), Journal of Computational and Graphical Statistics 26(1):1--13, proposed fitting several multivariate count distributions by iteratively reweighted Poisson regressions (IRPR). This is because the expected information matrices are expensive to compute. The VGAM R package can, in theory, be adapted to perform IRPR because its main algorithm is iteratively reweighted least squares. This project is to accomplish this; some VGAM family functions need to be written by adapting poissonff(). This project would suit a student with a good mark in STATS 310 and 330, as well as knowing R well (STATS 782because VGAM uses S4 object-oriented programming features).

Requirements: STATS 310,330 and 782

Contact: Thomas Yee (**t.yee@auckland.ac.nz**)

Fixed-effects vs mixed-effects models for grouped binomial data

This research seeks to estimate a population-wide (marginal) relationship between binomial data and a univariate covariate. The data are grouped with substantial-between group variability. Moreover, the groups from which the binomial data are measured can vary massively in size, although the sample size from each group remains roughly the same. This scenario is encountered in research on the size-selectivity of fishing gear, with some researchers choosing to use over-dispersed fixed-effects models, and others using mixed-effects models. Both have their weaknesses. This research will likely investigate a mixed-effects model that reweights predictions according to group size. Simulation will be essential under a variety of scenarios, and inclusion into R packages SELECT and selfisher. An abundance of data is available.

Requirements: Excellent grade in STATS 730. Strong R programming skills would be an advantage.

Contact: Russell Millar (r.millar@auckland.ac.nz)

A better threshold for Cook's distance

STATS20x uses a Cook's D threshold of 0.4 to label an observation as influential, but this is too low a threshold for small datasets and too high a threshold for large datasets. The current literature does not provide any consistent guidance. This project will use simulation to suggest a better threshold that depends on the number of observations (and possibly the number of parameters). The issue of false discover rate (FDR) may also be a consideration.

Requirements: Excellent grades in STATS 210 and 330. Strong R programming skills would be an advantage.

Contact: Russell Millar (r.millar@auckland.ac.nz)

Monotone splines for binomial data

This research will investigate the use of monotone splines for fitting to binomial data when it can safely be assumed that the relationship with the explanatory variable is monotone. This will require implementation of code that largely can already be found online. It will need to be customized for inclusion in the R package SELECT (used for fitting retention curves to fish catch data) in the situation where probability of retention increases with fish length. Methods to estimate L50 (length of 50% retention probability), and simulation comparison with parametric curves will be required. An abundance of data is available.

Requirements: Excellent grades in STATS 310 and 330. Strong R programming skills would be an advantage.

Contact: Russell Millar (r.millar@auckland.ac.nz)

Reliability of the Tweedie distribution

Over the last two or three years the Tweedie distribution has entered the statistical mainstream and some software now offer it as an option for the distribution of the data. The Tweedie family of distributions allows for automatic zero inflation of continuous data and has great potential in areas such as ecology. This research will look at the reliability and stability of the Tweedie distribution under a variety of scenarios with comparison to other more established distributions. An abundance of data is available.

Requirements: Excellent grades in STATS 310 and 330. Strong R programming skills would be an advantage.

Contact: Russell Millar (r.millar@auckland.ac.nz)

Splines versus model-averaged polynomials for fitting smooth curves to binomial data

In the literature, some authors use model-averaged polynomials (up to 4th order) to fit smooth curves to binomial data, while others prefer to use splines. This research will largely be simulation based and will explore the relative performance of these two methods under a variety of scenarios. The simulation context will be application to retention curves of fishing gears whereby retention probability is smoothly related to fish size.

Requirements: Excellent grades in STATS 210 and 330. Strong R programming skills would be an advantage.

Contact: Russell Millar (r.millar@auckland.ac.nz)

Exploring the association between alcohol availability and related illnesses

This project will explore the association between alcohol off-licenses in the Auckland region and the rate of alcohol-related hospitalisations. Ideally using the IDI, this project would explore scenarios that respond to different policy changes – such as the addition of more stores and/or other local government policy changes.

Requirements: This project requires somebody with strong R programming and data visualization skills. Good marks in STATS782 and/or 787 would be ideal. An interest in GIScience would be useful

Contact: Dan Exeter (d.exeter@auckland.ac.nz)

An on-demand community profile tool

Despite its limitations, Census data remains a reasonably accessible data source for community groups to obtain population statistics about their neighbourhoods. Most 'profiles' that are available at present are for Auckand Council Local Boards or Wards, Electorates, or District Health Boards. This project will involve collating a set of data from the 2018 Census to generate profiles in the form of infographics and geospatial maps.

Requirements: This project requires somebody with strong R programming and data visualization skills. Good marks in STATS782 and/or 787 would be ideal. An interest in GIScience, population health would be useful

Contact: Dan Exeter (d.exeter@auckland.ac.nz)

Exploring the association between deprivation and wellbeing among older people in Aotearoa

Opportunities are available for Hons or Masters students to explore associations between different health outcomes and the measurement of deprivation among the population aged 65+. We have previously developed methods using the 2013 Census, which we wish to explore further. Topics include using existing tools to explore measures such as comorbidities, post-operative recovery, cardiovascular disease and dementia. Alternatively, there is an opportunity to explore methods of multiple imputation to improve the completeness and coverage of some variables related to social position of older people.

Requirements: This project requires somebody with strong R programming with good data management and statistical modelling skills. Good marks in STATS769, 782 and/or 787 would be ideal. An interest in GIScience, population health would be useful

Contact: Dan Exeter (d.exeter@auckland.ac.nz)

Psychosocial well-being of people living in aged residential care

The objective of this study it to investigate the impacts of the COVID-19 pandemic, which included two nationwide lockdowns, on the health and psychosocial well-being of Māori, Pacific Peoples, Asians and New Zealand Europeans living in aged residential care. interRAI Long-Term Care Facilities (interRAI LTCF) is an internationally developed comprehensive geriatric assessment which provides information on 250 demographic, clinical and psychosocial factors.
In this study, interRAI assessments of Māori, Pacific Peoples, Asians and New Zealand Europeans (aged 60 years and over) completed in 2021 will be compared with interRAI assessments of the same ethnicities in 2019 and 2020. Physical, cognitive, psychosocial and service utilisation indicators will be included in the analyses and modellin

Requirements: Good grade in STATS380 and STATS330 or equivalents

Contact: Claudia Rivera-Rodriguez (c.rodriguez@auckland.ac.nz)

Trends on academic outcomes from Statistics students

This project aims to analyse longitudinal data on academic outcomes in order to identify patterns and trends over time. The project aims to evaluate whether these trends are the same for certain groups such as ethnic groups, ESOL students, among others.

The data is on academic outcomes from STATS20X and STATS10X courses over 10 years.

Requirements: The project requires R programming (STATS380 or equivalent) and modelling (STATS330 or equivalent).

Contact: Claudia Rivera-Rodriguez (c.rodriguez@auckland.ac.nz), Stephanie Budgett (s.budgett@auckland.ac.nz)