Postgraduate Project List for 2021

Machine Learning/Predictive Analytics with R

This project will involve working with the iNZight development team on developing capabilities for building predictive models using automated ensemble tools such as TidyModels and TensorFlow and simpler more understandable tools such as regression models and classification & regression trees. This is a good project for building data-science skills that should suit students with interests in computing and statistics. It is also an opportunity to learn valuable skills from the experienced members of the iNZight team.

Requirements: Very good grades in Stats 310 and 330, or 369 or equivalent and good R-programming skills.

Supervisor: Chris Wild, c.wild@auckland.ac.nz

———

Graphics for Cluster Analysis/Unsupervised learning and other graphics for multivariate-Y data

Unsupervised learning, known as Cluster Analysis or Clustering in Statistics, has the objective of grouping a set of objects (based on the data we have on them) in such a way that objects in the same group/cluster are more similar to one another in some sense than they are to members of other groups/clusters. This project has a particular emphasis on graphic representations of clustering and other multivariate graphics. A good project for building data-science skills that should suit students with interests in computing and statistics. An opportunity to learn valuable skills from the experienced members of the iNZight team.

Requirements: Very good grade in Stats 302 or equivalent and good R-programming skills.

Supervisor: Chris Wild, c.wild@auckland.ac.nz

———

{d3tourr}: An HTML widget for visualising high-dimensional data using tour

Tours are commonly-used visualisation methods for exploring high-dimensional data with projections. The R package {tourr} implements different tour algorithms, and dynamically renders the projected data via an X11 device. In this project, we will combine the R computing power with web technology for dynamic visualisation of multivariate data. We will develop an HTML widget using d3.js for R to make tour graphics faster and more accessible.

Requirements: Good programming skills in R and d3.js

Contact: Earo Wang (earo.wang@auckland.ac.nz)

———

Analysis of Covid-19 Time Series Data

Abstract: The student should collect the data from internet, do the pre-processing and perform the analysis (mainly in the frequency domain). The supervisor will provide explanations about the methods that should be applied as well as their implementations in Matlab. It is likely that some R-packages will be also used.

Requirements: Good programming skills and good mathematics. Prior knowledge on time series analysis is an advantage.

Supervisor: Ciprian Doru Giurcaneanu (c.giurcaneanu@auckland.ac.nz)

———

Psychometric properties of the Severity of Violence Against Women Scales in wāhine Māori

Co-supervised by Professor Denise Wilson, Director of the Taupua Waiora Centre for Māori Health Research, and Professor Jane Koziol-McLain, Co-Director of the Interdisciplinary Trauma Research Centre, both at Auckland University of Technology.

The SVAWS Total score and its three subscores (Marshall, 1992) have been validated in US populations, but not amongst Māori women. The purpose of this project is to assess the extent to which the suggested structure of the SVAWS (three scales subdivided into 9 dimensions) holds amongst Māori women using confirmatory factor analysis (CFA). The scale was administered at baseline and post-intervention during the isafe trial (Koziol-McClain, 2018). The specific objectives of the CFA will be to assess configural, metric and structural invariance, or lack thereof, of the SVAWS tool between Māori and non-Māori women, accounting appropriately for the interventional and longitudinal aspects of the data.

Prerequisite: knowledge of R, SAS, Amos, or some CFA or structural equations modelling software.
Desirable: some knowledge of the lavaan R package, PROC CALIS in SAS or Amos.
References

Koziol-McLain J, Vandal AC, Wilson D, Nada-Raja S, Dobbs T, McLean C, Sisk R, Eden KB & Glass NE. (2018). Efficacy of a Web-based safety decision aid for women experiencing intimate partner violence: a randomized controlled trial. *Journal of Medical Internet Research* **19**(12), e426. https://doi.org/10.2196/jmir.8617

Marshall, L. L. (1992). Development of the Severity of Violence Against Women Scales. *Journal of Family Violence* **7**(2), 103–121.

Supervisor: Alain Vandal (alain.vandal@auckland.ac.nz)

———

Hybridisation of two abstract scoring tools

Co-supervised by Dr Michael Meyer, Head of the Neonatal Intensive Care Unit, Middlemore Hospital.

Selection of submitted abstracts for presentation at a conference can be done using existing tools. Two such tools, "old" and "new", were assessed and found not to perform satisfactorily if the gold standard to be predicted is whether the research underlying the abstract is eventually published in article form in a peer-reviewed journal.

Two hundred abstracts, 100 published and 100 unpublished, were scored by four assessors: two using the "old" tool and two using the "new" tool. The goal of this work is to obtain a scoring tool by selecting the best items from either tool and create a hybridised abstract-scoring tool with a better performance in terms of predictive accuracy. Aside from a simple score, further investigations will involve use of classification and regression training to potentially improve the scoring, with the eventual creation of a Shiny app for use by conference Scientific Committee members.

Special features to overcome or leverage are the multiplicity of predictors and potentially different multiple predictions obtained for each abstract.

Prerequisite: knowledge of R programming
Desirable: knowledge of R packages glmnet, caret, and basics of machine learning in general

Supervisor: Alain Vandal (alain.vandal@auckland.ac.nz)

———

Statistics Education: Markov processes – a visual approach for learners

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

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 explore what new understandings may emerge for students as they learn about Markov processes using 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.

Supervisors: Azam Asanjarani (azam.asanjarani@auckland.ac.nz)andStephanie Budgett (s.budgett@auckland.ac.nz)

———

Title: Inference for Queues with Approximate Bayesian Computation

Abstract: Queueing systems can typically be modelled analytically in simple cases, but more realistic assumptions tend to require simulations. However, that makes it difficult to compare the models to data and infer the values of the queueing system's parameters. In this project you will develop ways of inferring queueing parameters from data using "Approximate Bayesian Computation", which essentially involves simulating datasets until you find ones that match the observed dataset.

Prerequisites: STATS 320, STATS 331, and excellent programming skills.

Supervisors: Brendon Brewer (bj.brewer@auckland.ac.nz), Azam Asanjarani (azam.asanjarani@auckland.ac.nz), and Ilze Ziedins (i.ziedins@auckland.ac.nz)

———

Title: Adaptive control of stochastic queueing networks

Abstract: The evolution of queuing systems often happens randomly and key variables/parameters of the system may be unknown or partially observed. Providing a stochastic model for these systems with the aim of improving the efficiency or forecasting and bringing them under on-line control, lead to reducing the customer waiting times, better server utilizations and stability. The main idea of this project is devising an appropriate and optimal model for a network of queues which 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.

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

———

Title: Stochastic modelling of patient's trajectories in a hospital

Abstract: The aim of this project is constructing a stochastic model for prediction of individual's progression through various stages of a disease, or provide a sensible contribution to build a prediction model that predicts the risks and chances of the expected trajectories of patients through intensive care units.

Requirements: the student needs to be familiar with stochastic processes such as Markov chains and queueing systems. Programming skills would be an advantage.

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

———

Title: Parameter and state estimation for queuing systems

Abstract: Queueing systems come up in a variety of situations in the real world. For instance, machines at a factory, telephone lines, traffic lights, runways at an airport, cashiers in a supermarket or doctors may not immediately supply their customers with the amount or kind of service they required. The evolution of queuing systems often happens randomly. To aid in understanding queuing systems, mathematical queueing models have been studied and employed for over a century. However, there are a few papers dealing with parameter and state estimation of queuing systems. The aim of this project is filling the literature gap in this area.

Requirements: The student will require to be familiar with queuing theory. A good grade in Stats 225 and/or Stats 320 or equivalent would be an advantage.

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

———

Title: Moment matching problem for truncated multivariate distributions

Abstract: The problem of matching distributional parameters to obtain desired moments is an interesting classic problem in the field of statistics and econometrics. The use of truncated distributions arises often in a wide variety of scientific problems. The aim of this project is to solve the moment matching problem using a novel dynamic method, specifically suited for truncated multivariate distributions.

Requirements: Mathematics skills (eg. proofs, limits) is essential. Some knowledge of basic probability and stochastic processes recommended (Stats 125, Stats325, Stats320).

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

———

Title: Scheduling for a processor sharing system

Abstract: In a variety of real-life queueing systems such as in manufacture, 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 kind of service they required. In these situations, for reducing the waiting time in the queue and act fairly to each request, we apply scheduling policies to determine which requests in the queue are serviced at any point in time, how much time is spent on each, and what happens when a new request arrives. The aim of this project is solving the problem of scheduling arrivals to a congestion system (such as traffic intersection) with a finite number of users having identical deterministic demand sizes.

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

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

———

When can Taylor's variance power law be applied to real data?

Taylor's power law (https://en.wikipedia.org/wiki/Taylor%27s_law) says that the variance of a count (or other measure of abundance, such as weight) is of the form var=a*mu^b, where mu is the expected value and a and b are parameters to be estimated.

General purpose model fitting software (such as Stan or TMB) enable Taylor's power law to be implemented for a wide range of distributions (e.g., negative binomial and lognormal) and classes of models. The question of interest is whether it provides a more parsimonious fit, or does it make the model overlying complex or computationally unstable.

Using appropriate model performance criteria, this work will re-analyse existing datasets and perform simulation studies to determine when Taylor's law can safely be applied.

This topic will require a good grade in STATS 730 or 731.

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

———

Develop an R-shiny tool to allow to calculate site-adjusted water quality guidelines

Co-supervised by Dr Jennifer Gadd, head of the Urban Aquatic Environments group at NIWA.

The water quality guidelines for zinc in freshwater are based on a species sensitivity distribution (SSD), which has a probabilistic model, such as log-normal, log-logistic or Burr Type III, fitted to calculate guidelines (at varying percentile levels). The method for zinc first requires the individual data values for the SSD to be adjusted to a standard water quality, either default values, or to user-added values. These are three multiple linear regression models used for this (one each for fish, invertebrates and algae) to account for the fact that zinc is less toxic in the presence of dissolved organic carbon (DOC), calcium and magnesium (hardness) and at low pH. An R shiny tool is required that would allow people to calculate the guideline at whatever pH, hardness and DOC they want, based on what they measure in their water. The data need to be adjusted based on each MLR model, then the SSD generated, and guidelines calculated.

The student should have good R programming skills but is not expected to have come across RShiny; a creative streak would be an advantage.

Contact: Charlotte Jones-Todd c.jonestodd@auckland.ac.nz

I often have other projects available, please take a look at https://cmjt.github.io/ and if any interest you feel free to contact me.

———

Estimating parameters of a void process using a Bayesian approach

Spatial point processes are mathematical models that describe the arrangement of objects, or events, located in space and/or time. The patterns formed by these points (objects or events) are analysed and offer insights into the underlying processes creating the observed patterns. Patterns observed may contain clusters of points or areas (voids) devoid of any points. This project will involve using a Bayesian approach to estimate parameters of a void point process.

No prior knowledge of point processes is required; however, a student should be comfortable with the concepts covered in STATS730 and STATS731.

Contact: Charlotte Jones-Todd c.jonestodd@auckland.ac.nz

I often have other projects available, please take a look at https://cmjt.github.io/ and if any interest you feel free to contact me.

———

What happens if we ignore self-excitement in point process models?

Events cluster in time and space: tweets go viral, flurries of nearby earthquakes occur in quick succession. Temporal and spatial proximity are major factors in the chain reaction of events. However, how and why these spatiotemporal clusters form is much more complex: the mechanisms that generate clusters of events are often self-exciting.

This project will look at what happens when self-excitement is ignored. No familiarity with point processes is needed; however, strong R programming skills are required and familiarity with C++ would be advantageous. Good grades in STATS310 and STATS330 would also be an advantage.

Contact: Charlotte Jones-Todd c.jonestodd@auckland.ac.nz

I often have other projects available, please take a look at https://cmjt.github.io/ and if any interest you feel free to contact me.

———

The role of the 'mesh' when using the INLA approach to fit a point process model

Spatial point processes are mathematical models that describe the arrangement of objects, or events, located in space and/or time. The patterns formed by these points (objects or events) are analysed and offer insights into the underlying processes creating the observed patterns. A log-Gaussian Cox process (LGCP) is a type of point process typically used to model clustered point patterns. To fit such a model the Integrated Nested Laplace Approximation (INLA) approach may be taken, which relies on the construction of a Delauney triangulation (mesh) as part of the model fitting procedure. This project will investigate how different properties of this mesh affect parameter estimation and its role when fitting LGCPs.

No familiarity with point processes is needed; however, strong R programming skills are required. A student should also be comfortable with the concepts covered in STATS730 and STATS731.

Contact: Charlotte Jones-Todd c.jonestodd@auckland.ac.nz

I often have other projects available, please take a look at https://cmjt.github.io/ and if any interest you feel free to contact me.

———

Project title:"**Stochastic models for biodiversity"**

This project will consider some stochastic models for biodiversity. In particular, some neutral models for extinctions and speciations using branching processes will be investigated mathematically, including the genealogical structure of reconstructed phylogenetic trees. 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: Basic probability (eg. Stats125) and some mathematics (eg. proofs, limits) is essential. Some knowledge of stochastic processes recommended (eg. Stats325, Stats320).

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

———

Project title: "**Mathematical Population Genetics"**
Supervisor: **Simon Harris**

This project will present an introduction to mathematical population genetics and coalescent processes, including the Kingman coalescent. Coalescent processes provide models for the genealogies (family trees) that are constructed backwards in time from samples of the present population, with ancestral lineages merging together whenever they first share a common ancestry. The student will follow a course of directed reading covering the necessary background material in probability (such as Markov chains), the Kingman coalescent, some related models, and some applications.

Requirements: Basic probability (eg. Stats125) and some mathematics (eg. proofs, limits) is essential. Some knowledge of stochastic processes recommended (eg. Stats325, Stats320).

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

———

Project title: "**Inhomogeneous branching Brownian motions"**
Supervisor: **Simon Harris**

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: Basic probability (eg. Stats125) and good mathematics (eg. proofs, limits, calculus, differential equations) is essential. Good knowledge of stochastic processes or Markov chains also strongly recommended (eg. Stats325, Stats320).

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

———

Comparing and averaging imperfect models in the Bayesian inference

Dealing with uncertainty in model choice and averaging is one of the fundamental problem in statistics. Model evidence (marginal likelihood) is often numerically approximated and numerous estimators have been proposed e.g., bridge sampling, Chibs method, importance sampling. It is still remained to be difficult and often expensive to compute in practice. There is a rich literature on alternative methods for weighting models. For example, the Bayesian information criterion is used to weight models. In this project, we will investigate the divergence based weighting method and examine it for various problems. Ideally, a student working on this project would have strong computational skills of R and a solid understanding of Bayesian inference and statistical methods (e.g., STATS 331 or 731 or 210 or 310).

Contact : Kate Lee (kate.lee@auckland.ac.nz)

———

Catching up with R graphics

The 'grImport' and 'grImport2' packages for R import PostScript and SVG images so that they can be drawn using 'grid' in R and the 'gridSVG' package for R exports plots drawn with the 'grid' package to an SVG format. The 'grid' package has recently added support for gradients, patterns, masks, and clipping paths, so the 'grImport' and 'grImport2' and 'gridSVG' packages all need to be updated to handle these new features.

Requirements:
Excellent R programming skills and comfortable with Linux. A familiarity with SVG would be helpful.

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

———

Statistics Education: Joint and conditional probability statements

This project would suit someone with a genuine interest in Statistics education. It would include conducting a literature review on how students develop an understanding of both joint probability and conditional probability. This means that skills in reading 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.

Supervisors:

Stephanie Budgett (s.budgett@auckland.ac.nz) and Marie Fitch (m.fitch@auckland.ac.nz)

———

Automating visual analysis for time series data

Time series data often carries rich information, including trend, seasonality, and outliers. A variety of statistical tools have been developed for mining these temporal characteristics. This work presents an automated and interactive dashboard to generate an ensemble of linked graphics for user-supplied time series of any time resolutions, augmented with graphical interpretations. These preliminary results will be in turn used to recommend a statistical model in order to forecast series of interests.

This project aims to lower the barrier for time series analysis. You will need to develop a performant and aesthetic R Shiny dashboard for the project.

- Reference: "Forecasting: Principles and Practice" (available online: https://otexts.com/fpp3/)
- Requirements: A good knowledge of time series analysis, and R Shiny programming

Contact: Earo Wang (earo.wang@auckland.ac.nz)

———

The Piranha Problem

This is a problem in understanding applied statistics, named by Andrew Gelman.

Small research studies in, eg, psychology or diet often produce quite striking results from quite minor interventions. Gelman argues that these can't actually be true in any useful sense, because you would be able to combine the interventions to get really huge effects, and we don't believe this is possible. Either there are lots of interactions, so that even the direction of the effect of one intervention depends on the others, or the results are wrong. The metaphor is of a large collection of piranha fish in a small tank, where the biggest one would just eat the others.

The project is to quantify this analytically and by simulation in some specific cases. Yes, it's a bit vague. Students should have a good understanding of linear and logistic regression (STATS 330), and some programming ability in R.

Supervisor: Thomas Lumley (ts.lumley@auckland.ac.nz)

———

Comparing tests for survey data

When fitting generalised linear models to data from multistage surveys there are even more choices of test than for ordinary data. There are score and Wald tests that combine different parameters based on their sample importance, and score, Wald, and likelihood ratio tests that combine different parameters based on their population importance. There are also different approximations to the asymptotic distribution of these tests. We have almost no idea of the relativel operating characteristics of these tests.

The project would compare the extent to which these tests achieve their nominal size, and their power against various alternatives, in a range of simulated multistage surveys, and also illustrate the results in actual surveys. This should be a relatively straightforward project if you have taken STATS 740 and also have some ability in programming.

Supervisor: Thomas Lumley (ts.lumley@auckland.ac.nz)

———

Score-based inference for multiple imputation

Multiple imputation handles missing data by filling in a distribution of multiple plausible values. If it is done right, and under only moderately unreasonable assumptions, multiple imputation can provide valid inference in the presence of missing data: consistent parameter estimates and valid uncertainty estimates.

The most common approach to inference after multiple imputation is due to Rubin and uses an argument motivated by Bayesian considerations. Unfortunately, the argument assumes that the imputation model and the model being used for analysis are 'the same' in quite a strong sense, which is often not true.

Robins & Wang proposed an approach to inference that does not make these assumptions, and that was designed to be easy to implement on a computer. They aren't programmers, so they weren't entirely right about the ease of implementation, but it is feasible. The project is to identify a reasonably general set of imputation and analysis models that can be implemented in a fairly automated way, and to implement it. You'll need 310 and 330 as background, and more maths will help. You will also need to be interested in statistical computing as a design problem, and something like 782 would be helpful.

It might be possible to split this up into projects for two people who are willing to collaborate to some extent.

Supervisor: Thomas Lumley (ts.lumley@auckland.ac.nz)

———

Various Projects

I am open to discussions with students interested in Bayesian Inference, Information Theory, or decentralised publication, who would like to do a project with me. I have various ideas that are not sufficiently well-formed to get an abstract written here, but which could crystallise into a real projcet.

Supervisor: Brendon Brewer (bj.brewer@auckland.ac.nz)

———

Title: Masters- Understanding behavior change in a complex participatory community intervention to improve maternal and child survival in rural Malawi: self-esteem, empowerment, co-coverage and equity

Abstract: A large cluster randomized trial of a participatory community intervention with women's groups was conducted in rural district of Malawi (https://www.ncbi.nlm.nih.gov/pubmed/23683639). This intervention demonstrated an impact on maternal and child mortality, but the mechanism of behaviour change is still not understood. This project will develop the trial analysis by exploring two additional areas: 1) the impact of intervention participation on the "co-coverage" of key maternal, newborn and child health behaviours; and 2) the mediating role of empowerment and self-esteem. There is growing literature on the measurement of "co-coverage" for complex maternal and child health interventions1. This approach acknowledges that improvements in health and reductions in mortality due to complex public health interventions may arise from the adoption of multiple health behaviours. Indeed, public health interventions that target multiple behaviours simultaneously may have the most powerful effects. However, the analysis of trials typically focuses on the impact of interventions on separate individual behaviours. This project will involve a secondary analysis of data from the trial conducted in Malawi. The student will create a "co-coverage" index relevant to this study, and use this to measure the impact of the intervention on health behaviour. Further analyses will explore other ways of grouping health behavior variables, such as calculation of a composite index.

There is a well-established link between self-esteem and mental health, and evidence for the role of empowerment in family planning and maternal health2-4, but less is known about how self-esteem and empowerment may impact upon uptake of health services and childcare behaviours. The student will review the literature on measures of self-esteem and empowerment, and the mediating role of these on health behaviours. They will construct a self-esteem/empowerment scale using the questionnaire data collected, and use this to build a multivariable model looking at the relationship between group membership and health behaviour, then explore how self-esteem mediates this effect.

The specific objectives of the study are to:

Review literature on co-coverage and composite indices and develop a co-coverage index

Review literature on empowerment, self-esteem and maternal and child health

Build multivariable models to explore the impact of intervention participation on co-coverage and composite indices

Conduct equity analyses to explore the distribution of co-coverage by household wealth, maternal education and intervention participation

Evaluate the 'equity impact' of the intervention on co-coverage by household wealth

Explore the study data on empowerment and self-esteem and use grouping methods, such as principal components analysis or factor analysis to important dimensions of self-esteem

Use these findings to inform the development of an empowerment/self-esteem scale

Build a multivariable model to explore the impact of intervention participation on empowerment/self-esteem

Build a multivariable model to explore the impact of empowerment/self-esteem on health behaviours

Build a multivariable model to explore the mediating role of empowerment/self-esteem in the relationship between intervention participation and health behaviour

Write a research paper

Requirements

R programming, modelling and principal components analysis/factor analysis.

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

———

Title: Masters- Using big data to estimate the economic costs of dementia in New Zealand

Abstract: Dementia describes a cluster of symptoms that include memory loss, difficulties with thinking, problem solving or language, and functional impairment. Dementia can be caused by a number of neurodegenerative diseases, such as Alzheimer's disease and cerebrovascular disease. Currently in New Zealand, most of the systematically collected and detailed information on dementia is obtained through a suite of interRAI assessments, including the Home Care, Contact Assessment and Long-Term Care Facility versions. These versions of interRAI are standardised comprehensive geriatric assessments. Patients are referred to have an interRAI assessment by the Needs Assessment and Service Coordination (NASC) services after a series of screening processes. This study will focus on information collected by the

NASC services in Counties Manukau District Health Board (NASC-CMDHB) using the Home Care and Contact Assessment versions of interRAI completed between 1st July 2014 and 1st July 2019. This information will be linked to produce a large integrated dataset which will be used to: 1. Estimate the informal and formal costs of dementia care per patient; 2. Investigate the drivers of the costs of dementia including cultural and sociodemographic factors, severity of dementia, comorbidities and social support. Formal costs are medical care (primary care, secondary care, district nursing care) and social care (home care, day care, respite care and long-term residential care). Informal care includes care provided by unpaid (informal) carers, usually family members. The statistical methods used will be advanced design-based survey methods for prevalence and generalized estimating equations for regression models and correlated/longitudinal data. Previous estimates of the economic costs of dementia in New Zealand were based on international studies where their populations and health and social care systems are different. This new local knowledge will have implications for estimating the socio-economic impact of dementia in New Zealand.

Requirements

R programming, modelling and survey sampling theory.

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

———

Modelling the scale of spatial interactions between key sandflat species

Inter-species and species-habitat interactions are usually modelled as if they occur at a fixed scale, however, the strength of any interaction is likely to be a function of the scale at which different species experience their environment and also the scale at which observations are made. Aspects of the spatial scale that affect observations include grain (the size of the sampling unit), lag (the distance between samples) and extent (the area over which the observations are made). This study will assess the interactions observed between key sandflat species in different sized windows of observation by exploring the effect of varying lag and extent on simple correlations and analysing cross-correlograms.

Prerequisites: r programming to extract different sized windows, basic GIS ability, at least one biology course.

Contact Prof. Judi Hewitt judi.hewitt@niwa.co.nz

———

Simulating non-random community assemblage processes

Present meta-community theory suggests that the factors controlling how species are built vary from environmental filtering, source-sink dynamics (dispersal) and species interactions to stochastic demographic processes. This project would integrate the use of biological traits that represent dispersal capability, environmental selection and adult-juvenile-adult interactions into community assemblage theory. Monte Carlo simulations driven by biological traits would be used to build communities. The Metacom package would be used to assess whether the communities can be appropriately assigned to particular species assemblage mechanisms.

Prerequisites: understanding of Monte Carlo simulations, at least one biology course.

Contact Prof. Judi Hewitt judi.hewitt@niwa.co.nz

———

Investigating the efficiency of multivariate species distribution models to represent biodiversity

Corelative models (e.g. GLM, GAM, Boosted Regression Tree model, Random Forest model) are routinely used to link biodiversity observations at specific sites to the prevailing environmental conditions at those sites. Once biodiversity-environment relationships have been quantified these can be used to make predictions in space and in time by projecting the model onto available environmental layers (termed Species Distribution Models (SDM)). A recently developed multivariate extension to Random Forest modelling called Gradient Forest modelling predicts species turnover (the differentiation of species over space) rather than individual species' distributions. It is hypothesised that Gradient Forest models may better represent biological communities when sampling is limited compared to individual SDMs where taxonomic surrogacy would have to be assumed, i.e. that one species captures variation in another. This project would compare the ability of Gradient Forest models and individual SDMs to represent demersal fish communities under different sampling regime scenarios using a large biological dataset (> 60, 000 trawls targeting demersal fish) and high-resolution environmental predictor variables (1km grid resolution) across the New Zealand Exclusive Economic Zone.

Prerequisites: r programming (correlative models, GLM, GAM, BRT, RF; stratified and random sampling), basic GIS ability.

Contact Prof. Judi Hewitt judi.hewitt@niwa.co.nz

———

Detecting changes in species co-occurrences along environmental gradients.

This project merges the use of pairwise species co-occurrence analysis, fuzzy clustering and indicator values derived from specificity and fidelity of group membership to unpack changes occurring in benthic macrofaunal communities along environmental gradients. Clustering techniques are commonly used to determine how sites are clustered based on the species that live there. Hierarchical clustering can be used to break down groups of sites at different levels of similarities, while fuzzy clustering allows a site to move from one group to another and to belong to multiple groups with varying levels of confidence. Specificity and fidelity of the species comprising the hierarchical clusters are used to create indicator species or their associated environmental drivers. Relating the degree of specificity and fidelity of indicator species and environmental variables to degree of fuzziness creates an estimate of uncertainty that could be compared to the probability of indicator species co-occurrences.

Prerequisites: r programming for multivariate analysis

Contact Prof. Judi Hewitt judi.hewitt@niwa.co.nz

———

Generally-Altered, -Inflated and -Truncated 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 and -Truncated Regression (GAIT regression) and implemented in the VGAM R package for three 1-parameter families. In GAIT regression the three operators apply to general sets rather than {0}. Also, the three operators may appear simultaneously in a single model.

Elements of the three 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 "GAIT Mix-MLM combo" model has five special value types. GAIT 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.

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)

———

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 (meyer@stat.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 (meyer@stat.auckland.ac.nz)

———

Hunting for Patterns in Andromeda's Globular Cluster Population

Big galaxies, like our own Milky Way, are orbited by small stellar systems known as globular clusters. These are old, and were witness to the birth of galaxies. But do they orbit randomly or in distinct populations? In a recent study, we found that the outer globular clusters of Andromeda, our nearest cosmological neighbour, orbited in two orthogonal populations, a result at odds with our ideas of galaxy evolution. In this study, we will explore the orbits of the inner globular clusters in a similar manner. Ideally, a student working on this project would have strong computational skills and a solid understanding of Bayesian inference.

Contact: Brendon Brewer (bj.brewer@auckland.ac.nz)

———

Nested Sampling for Experimental Design

In 2017, I developed a modified version of the Nested Sampling algorithm, which is usually used for Bayesian inference, so that it can compute quantities from information theory. In principle, this could be used as the foundation for doing Bayesian experimental design (how should an experiment be arranged so that we learn a lot about the quantities of interest?). The purpose of this project is to make this potential a reality and demonstrate it on some interesting example problems.

Ideally, a student working on this project would have strong computational skills and a solid understanding of Bayesian inference and Monte Carlo methods.

Contact: Brendon Brewer (bj.brewer@auckland.ac.nz)

———

Chronic Kidney Disease in a New Zealand population

Chronic kidney disease (CKD) is characterised by low glomerular filtration rate (GFR). CKD increases the risk of cardiovascular disease, (CVD) death and developing end stage renal disease (ESRD). Globally, CKD is now the 18th most common cause of death worldwide with a steady increase in disease burden of over 50% between 1990 and 2010 and a 50% increase in global years of life lost. The number of people with CKD in New Zealand (NZ) is currently unknown. An estimate of 7-10% for CKD stages 3 to 5 has been proposed. The number of patients requiring renal replacement therapy (RRT) for ESRD in New Zealand has increased steadily from 445 per million population (pmp) in 1995 including transplanted patients to 959 pmp in 2014, costing between $30,000 to $60,000 per patient per year amounting to 1-2% of NZ's entire healthcare expenditure5. The prevalence of ESRF is higher in Maori and Pacific people living in NZ, particularly in those with diabetes.

However, broader health impacts of CKD are not uniform. For some, the presence of CKD will be merely a diagnosis, with no appreciable impact on their life, while in others it will be a disease - strongly predictive of ESRD and need for ongoing dialysis and transplantation, and increased risk of CVD. Prediction models have the potential to stratify patients with CKD by their risk for these clinically important outcomes. Early detection of CKD in high risk groups is a health priority for primary care.

Prediction models have been developed for high risk populations, such as persons with CVD, or specific renal diseases such as diabetic or IgA nephropathy. Several population-based prediction equations have been published, although the most well validated are those developed with Canadian datasets – the Kidney Failure Risk Equation (KFRE). The KFRE has been validated in prediction datasets from 31 countries albeit with variable discrimination and calibration. Using this system, the expected 5-year incidence of various outcomes can be obtained by entering between 4 and 8 highly accessible and routinely collected demographic and biochemical patient characteristics into a web form and observing the output (http://www.qxmd.com/calculate-online/nephrology/kidneyfailure-risk-equation).

The New Zealand population is unique in terms of ethnic group composition and Māori and Pacific people have a disproportionate burden of ESRD. Little is known about the progression of CKD in New Zealand nor is there information about the burden of CKD pre-RRT.

Specific aims include:

1) Determine the prevalence of CKD at the time of their baseline examination in a working population of over 5,000 adults, and the prevalence of the diagnosis of CKD in this same population up to 40 years later.

2) Determine predictors of progressive CKD including the use of reno-toxic and renal-protective medications in the general population and by ethnicity where there are sufficient patients.

3) Explore the relationship between CKD and all-cause mortality and rate of hospitalisations in New Zealand patients as there is no data available from NZ.

4) Validate international risk scores for ESRD (calibration and discrimination) and to develop an ESRD risk prediction model based on demographic and clinical data.

5) Develop a risk prediction equation for risk of a first CVD event in patients with CKD.

Contact: Patricia Metcalf (p.metcalf@auckland.ac.nz)