Covid Wastewater Modelling

Supervisor: Thomas Lumley

Description: Fragments of the SARS-2-CoV viral genome can be found in wastewater (sewage). As testing and reporting of Covid decreases, these are a promising marker for Covid prevalence and may provide useful warnings about coming increases. The project involves working with a group of scientists coordinated at ESR on various modelling issues to do with forecasting Covid using wastewater data.

Requirements: a good grade in 330 or 762. Programming skills, and knowledge of time series and Bayesian inference would be advantageous.

How does rain affect buses?

Supervisor: Thomas Lumley

Description: This project involves modelling from Auckland Transport (and potentially Wellington and Christchurch) and from the Cliflo weather system to estimate the impact of weather on buses. You will need to obtain and store data using the AT and Cliflo APIs and clean and model it.

Requirements: STATS 369 and 380 or equivalent knowledge.

Bayesian sandwich estimator

Supervisor: Thomas Lumley

Description: There is a Bayesian analogue of the 'sandwich' variance estimator for linear models. The project is to implement it, using the 'brms' package in R.

Requirements: Good understanding of STATS 330, 331, and 310, and advanced programming skills.

Statistics on pairs

Supervisor: Thomas Lumley

Description: Some well-known statistics can be written either as a sum over observations or as a sum over pairs of observations. Familiar examples include the variance and the Mann-Whitney/ Wilcoxon test (which has different names in the two forms). When applied to clustered survey data, the natural ways of incorporating weights for individual observations or for pairs can give different statistics. This project will look (mostly by simulation) at how the two forms differ and how this depends on the design of the survey.

Requirements: No specific requirements, but this would be a good project for someone who is taking/has taken STATS 740 and is interested in theory

Group based trajectory models for high dimensional, modest sample size data

Supervisor: Beatrix Jones

Description: "Group based trajectory models," also known as "Latent class mixed models" are generally used for univariate or low dimensional data. This project will investigate the use of these models for time course data from the whole metabolome, in a crossover experiment of modest sample size. A clustering approach will be used to understand which metabolites have compatible latent classes, followed by joint modelling of those metabolites.

Requirements: A good understand of generalised linear models and/or mixed models. Existing R packages will be used, but student should be comfortable coding in R.

Resources:
lcmm: Extended Mixed Models Using Latent Classes and Latent Processes. https://cran.r-project.org/web/packages/lcmm/index.html

Joblin-Mills et al (2022) The impact of ethnicity and intra-pancreatic fat on the postprandial metabolome response to whey protein in overweight Asian Chinese and European Caucasian women with prediabetes. Frontiers in Clinical Diabetes and Healthcare 3. doi: 10.3389/fcdhc.2022.980856

Compositional data analysis to understand the association between ground cover flora and the soil microbiome

Supervisor: Beatrix Jones

Description: This project will study the associations between vineyard ground cover and vineyard management practices, time of year, and soil microbiota. Both the ground-cover and microbiota measurements are compositional in nature, eg for the vineyard ground cover, the proportion of a 1m^2 area covered by each type of plant must add to 1. The project will use specialised multivariate techniques for this type of data.

Requirements: Good mark in STATS302 or equivalent knowledge, and good R coding skills.

Resources:
Leong et al (2020). Using compositional principal component analysis to describe children's gut microbiota in relation to diet and body composition. American Journal of Clinical Nutrition, 111(1), 70-78. doi: 10.1093/ajcn/nqz270

Giraldo-Perez, P., Raw, V., Greven, M., Goddard, M.R., (2021) A small effect of conservation agriculture on soil biodiversity that differs between biological kingdoms and geographic locations, ISCIENCE doi: https://doi.org/10.1016/j.isci.2021.102280.

Analysis of food purchase panel data

Supervisor: Beatrix Jones

Description: The Nielsen IQ data is panel data (roughly the same households over time, although they can enter/leave the sample). This particular dataset has, over a few years, all food purchased to be prepared or consumed at home. The exact item purchased, and what was paid for it is recorded, as well as a classification of the item into one of about 300 categories. This data could be used to address many questions using conventional multivariate approaches (eg PCA, LDA). How does seasonality affect food purchases? How does the presence of children in the household affect purchases? What factors affect the cost of 'healthy food'?

Requirements: Good mark in STATS302 or equivalent knowledge, and good R coding skills.

Resources: Tawfiq E, Bradbury KE, Ni Mhurchu C. Does the prevalence of promotions on foods and beverages vary by product healthiness? A population-based study of household food and drink purchases in New Zealand. Public Health Nutr. 2021 Dec 20:1-9. doi: 10.1017/S1368980021004936.

Temporal model for extreme temperature in NZ

Supervisor: Kate Lee

Description: Extreme daily temperatures are known to be increased over years in NZ, like everywhere else. We will model the time varying extreme temperatures in the Bayesian framework. Variational inference will be also taken in to consideration for fast computing.

Requirements: Good programming skills in R and good understanding of Bayesian inference, MCMC techniques.Genealogies of samples from stochastic populations and biodiversity models

Genealogies of samples from stochastic populations and biodiversity models

Supervisor: Simon Harris

Description: 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).

Inhomogeneous branching Brownian motions

Supervisor: Simon Harris

Description: 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 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).

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

Supervisor: Thomas Yee

Description: 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).

VGAMviz

Supervisor: Thomas Yee

Description: 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.

Requirements: STATS 330, and STATS 380 or 782

Multinomial Logit Model

Supervisor: Thomas Yee

Description: The aim of this project is to improve the multinomial() family function in the VGAM R package. The multinomial logit model 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 write functions 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.

Multivariate count distributions estimated by iteratively reweighted Poisson regressions

Supervisor: Thomas Yee

Description: 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

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

Supervisor: Russell Millar

Description: 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.

A better threshold for Cook's distance

Supervisor: Russell Millar

Description: 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.

Monotone splines for binomial data

Supervisor: Russell Millar

Description: 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.

Reliability of the Tweedie distribution

Supervisor: Russell Millar

Description: 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.

Implementing the DVS Saliency Model in R

Supervisor: Paul Murrell

Description: Saliency models can be used to predict which parts of an image will attract visual attention. The Data Visualisation Saliency (DVS) Model is a saliency model that is designed specifically for predicting visual attention in data visualisations, so it is a useful tool for evaluating the effectiveness of data visualisations. Unfortunately, the DVS model has only been implemented in MATLAB code, so it is difficult and/or expensive to use. This project will look at reimplementing the DVS model in R code in order to make the DVS saliency model much more accessible to a wider audience. The final R code will be bundled in the form of an R package.

Requirements: Very good R coding skills. A good grade in least one of STATS 220, 380, and 782. This project would suit a student who likes to write code in their own spare time. A student who has a good grade in STATS 787 would be ideal.

Resources:
"Data Visualization Saliency Model: A Tool for Evaluating Abstract Data Visualizations", Matzen et al (2017).
https://www.osti.gov/pages/servlets/purl/1377597
The original MATLAB source code is also available from the supervisor.

Constraints: Software contributions must be licenced under the GNU General Public Licence.

CRAP Design Tools in R

Supervisor: Paul Murrell

Description: The CRAP design guidelines (Contrast, Repetition, Alignment, and Proximity) can be used to improve the effectiveness of data visualisations by placing emphasis on specific elements of a visualisation, clarifying and simplifying the structure of avisualisation, and connecting related components of a visualisation. This project will look at developing R functions that can assess a data visualisation in terms of the four CRAP dimensions. These R functions can then be used by the designer of a visualisation to assess the visualisation in terms of the CRAP guidelines. The final R code will be bundled in the form of an R package.

Requirements: Very good R coding skills. A good grade in least one of STATS 220, 380, and 782. This project would suit a student who likes to write code in their own spare time. A student who has a good grade in STATS 787 would be ideal.

Resources: "The Non-Designer's Design Book" by Robin Williams.

Constraints: Software contributions must be licensed under the GNU General Public Licence.

Enhancing the {hyperfun} package for R

Supervisor: Paul Murrell

Description: The HyperFun Project provides a language and interpreter for describing and viewing 3D scenes using Function Representation and Constructive Solid Geometry. The {hyperfun} package for R provides an R interface to the HyperFun language and interpreter, but it is an incomplete interface. This project would look at enhancing the {hyperfun} package to support more of the HyperFun language. For example, there are functions in the HyperFun FRep library that do not yet have an R interface. The HyperFun language also allows attributes such as colour and material properties to be specified, but these are not yet supported in {hyperfun}. Finally, there are potential extensions of HyperFun to explore, such as transformations like champfer and interpolation.

Requirements: Very good R coding skills. A good grade in least one of STATS 220, 380, and 782. This project would suit a student who likes to write code in their own spare time.

Resources:
http://paulbourke.net/dataformats/hyperfun/
https://github.com/pmur002/hyperfun/
https://www.stat.auckland.ac.nz/~paul/Talks/NZSA2022/

Constraints: Software contributions must be licenced under the Common Good Public Licence.

Associations between social deprivation and Venous Leg Ulcers

Supervisor: Yannan Jiang

Description: Venous leg ulcers (VLU) are the severest presentation of chronic venous insufficiency and a chronic relapsing remitting condition. VLU present as unhealing wounds of the lower leg, usually coupled with surrounding skins changes. Incidence of VLU increases with age. There has been very limited inquiry into associations between social deprivation and VLU; examination of a UK cohort suggested a socio-economic gradient existed (Petherick, Cullum, & Pickett 2013), but there has been no similar exploration in New Zealand. The 4VLU Collaboration has conducted five clinical trials in different centres in New Zealand involving 981 participants between 2003 and 2022. There is an opportunity to combine the trial datasets and describe the patient cohort using the New Zealand Deprivation Index (NZDep) and the Index of Multiple Deprivation (IMD) with the geocoding.

Co-supervisors: Associate Professor Daniel Exeter who developed the IMD (https://imdmap.auckland.ac.nz/download/), and Professor Andrew Jull who was principal investigator of the VLU trials.

Requirements: Competence in data linkage and statistical analysis using R and SAS; excellent writing and communication skills.

Modelling morphometric data of animal populations collected by drones

Supervisor: Ben Stevenson

Description: Measuring the size of animals by hand can be difficult because they might be big and scary (e.g., grizzly bears), react negatively to handling (e.g., manta rays), difficult to access (e.g., mountain goats), or too large for your tape measure (e.g., blue whales). We can avoid these problems by flying drones equipped with cameras over animals, but this method introduces the issue of measurement error: when we measure the size of an animal using a drone we don't get the answer quite right.

Myself and collaborators have developed a method to estimate population-level distributions of morphometric measurements in a way that accommodates drone measurement error. This project involves extending our existing model, and developing a user-friendly R implementation.

Requirements:
- A solid understanding of statistical theory (e.g., an excellent grade in STATS 310).
- Excellent R programming skills.
- Experience with C++, or a willingness to learn.

Fitting acoustic spatial capture-recapture models with `acre`

Supervisor: Ben Stevenson

Description: The R package `ascr` has been in development for over a decade, and provides functions to fit acoustic spatial capture-recapture (SCR) models. A new package, `acre` is currently under development, and is a complete rewrite of `ascr`, and includes a variety of new features.

In this project, we will refit models to data previously analysed using \texttt{ascr} and make extensions using `acre`'s new features. Our goals are (1) to test \texttt{acre} to ensure it gives comparable output to `ascr` when it is supposed to, (2) try out the new features of `acre` to ensure they work as we'd expect, and (3) create example analyses to use as training materials for ecologists.

Requirements:
- Excellent R programming skill (top grades in courses like STATS 220 and STATS 380 would be a big advantage).

Visualising Auckland's buses

Supervisor: Thomas Lumley

Description: Design and implement a visualisation of a day of Auckland's buses, showing the impact of delays and cancellations and full buses. Should run reproducibly given a day of stored data.

Requirements: good programming skills; you will likely need to learn some Javascript as well as R. STATS 369 and 380 would be helpful but are not required.

Multiple frame sampling for surveys

Supervisor: Thomas Lumley

Description: A key element of a survey is the sampling frame; the population list that people are (explicitly or implicitly) sampled from. Some surveys use two or more sampling frames, either for completeness or to oversample a group of people. Estimation in multiple-frame surveys is understood theoretically but not widely implemented. This project is to implement multiple-frame estimation for the survey package

Required: good R programming skills.
Recommended: STATS 740. STATS 762 or 763

Tree-based algorithms to predict the reproductive state of cereal crop pest snails in the southern states of Australia from environmental factors

Supervisors: Dr Kathy Ruggiero, Dr Lisa Chen

Description: Exports of Barley and wheat to China bring into the Australian economy an estimated AU$1.4 billion each year. Crop contamination by invasive snail species are a constant threat to this major Australian market potentially leading to devaluation of crops upon receipt and/or blocking of market access. Eliminating this crop pest is of critical economic importance.

The Australian Grains Research Development Corporation has, over the past decade, invested in a series of projects aimed at improving snail management, with a focus on baiting. Timing is an essential factor in baiting success, particularly when the goal is to prevent snails from reproducing since it is the small infant snails which are the major culprits of contamination.
This project focuses on data collected from a large-scale study of invasive snails conducted across several sites in each of the southern states of mainland Australia. The purpose of the study is to determine whether environmental factors such as meso- and micro-climate serve as good predictors of when snails are entering their reproductive state. This would then enable farmers to know when they should lay out their bait. In this project you will apply different tree-based algorithms to assess and compare the quality of the solution offered by each.

Requirements: Confident R user, at least a B+ grade in both STATS 330 and STATS 369. Successful completion of either STATS 240 or STATS 340 is preferred but not essential.

Resources:
Australia's agriculture exports to China, AgriInvestor, (https://www.agriinvestor.com/australias-agriculture-exports-to-china-in-numbers/
GRDC Update Papers (https://grdc.com.au/resources-and-publications/grdc-update-papers/tab-content/grdc-update-papers/2020/02/snail-management-learnings-from-recent-studies)

Classifying the reproductive state of crop pest snails based on albumen gland measurements.

Supervisor: Dr Kathy Ruggiero

Optimal efficacy of snail baits is achieved when their application coincides with the set of environmental conditions which trigger reproductive activity in adult snails. The length of an adult snail's albumen gland, which enlarges in preparation for reproduction, is considered an indicator of its reproductive state (i.e. active or inactive). Our data shows that the albumen gland lengths among the individuals collected in the monthly samples for our study vary considerably, showing that the population does not become reproductively active (and return to being inactive) in unison, but in waves across time. This is an important finding since, in order to identify the environmental variables predictive of reproductive state, we must first be able to label individuals as being in reproductive or non-reproductive state. In this project the student will: apply unsupervised classification methods to identify reproductively and inactive snails across time, explore the availability and use of R packages for performing unsupervised classification, and explore the use of resampling methods to estimate clustering validity.

Requirements: Confident R user, at least a B+ grade in both STATS 330 and STATS 369. Successful completion of either STATS 240 or STATS 340 is preferred but not essential.

Projects becoming available in Semester 2

Adaptive control of stochastic queueing networks

Supervisor: Azam Asanjarani

Description: 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.

Stochastic modelling of patient’s flow in a hospital

Supervisor: Azam Asanjarani

Description: 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.

Moment matching problem for truncated multivariate distributions

Supervisor: Azam Asanjarani

Description: 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) are essential. Some knowledge of basic probability and stochastic processes is recommended (STATS 125, STATS 325, STATS 320).

Scheduling for a processor sharing system

Supervisor: Azam Asanjarani

Description: 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.