Department of Statistical Sciences (DoSS)
2022 Summer Undergraduate Research

The Department of Statistical Sciences is seeking applicants interested in conducting a summer research project under faculty supervision. Projects will take place for a 16 week period from May-Aug, 2023. The following research awards are being offered this summer: NSERC USRAs ($6,000), UTEA ($7,500), DoSS SURAs ($2000) and the CIHR and SSHRC Black Student Undergraduate Student Research Awards ($7,500). In addition, Students also interested in doing a STA496H1/STA497H1 Independent Reading course during the Summer term should fill out this form as well.

All students who are currently enrolled in a Department of Statistical Sciences (DoSS) Specialist or Major program and who have completed STA302H1 with at least a B+ by the time of application are eligible to be considered for one of these awards. You will be asked to rank your top 3 projects. Students will be considered based on their academic record, experience and their statement of research interest. Students who are shortlisted for specific projects may be invited for an interview with the prospective supervisor.

Projects:

Spatial and Temporal Analysis of Emissions from Oil and Gas Development

Supervised by: Meredith Franklin

With the rise in unconventional oil and gas development (UOGD), domestic oil and gas production has increased in the United States and Canada to its highest level in over a decade. UOGD extraction techniques including horizontal drilling and high-volume hydraulic fracturing have enabled the exploitation of previously inaccessible or uneconomic shale plays, resulting in thousands of extraction sites across relatively small geographic areas. Subsequently, oil and gas extraction has become more common near where people live and work, increasing the potential for human exposure to air contaminants and noise.

Using data we are collecting from a multi-instrument site in the Permian Basin, New Mexico, we will quantify the magnitude, frequency and duration of UOGD chemicals at multiple temporal and spatial scales. By coupling our measurements with information on meteorology and land use we will characterize human exposures over the study regions in order to inform subsequent health studies. Finally, by leveraging satellite observations of thermal sources we will identify the locations of UOGD flaring in the study areas, and by combining them with time-resolved monitoring data we will be able to assess the chemical exposures related specifically to flaring.

Analyses will primarily include data fusion (from multiple databases), data visualization (maps), and time series modeling.

Risk Aware Reinforcement Learning

Supervised by: Sebastian Jaimungal

Reinforcement learning (RL) typically aims to minimize the expected sum of discounted costs over a set of policies, where the policy affects those costs. In many situations, however, expectation is too crude a measure of performance of a policy, and instead one may wish to assess left and right tail risks of policies.

This project will involve learning about risk aware RL and investigating its application to carbon capture and clean energy projects.

Computational Tools for Exploring Multimodal Posterior Distributions

Supervised by: Vianey Leos Barajas

In the field of ecological statistics, finite and dependent mixture models are often used to identify a set of K patterns in data that can serve as a proxies ecological processes of interest such as animal behavior, presence/absence and changing life stages. When working with large data sets from multiple individuals, we are tasked with both identifying the likely value of K as well as testing our assumptions of how to pool the parameters across individuals. Mixture models are notoriously difficult to identify in practice and oftentimes we are faced with exploring multimodal posterior distributions to conduct inference. While this can be viewed as a simple computational challenge, generally the multiple modes also reveal the manner in which the model is misspecified and uncovers unexplained heterogeneity across the individuals of interest.

In this project, we will explore parallel tempering techniques as well as newer algorithms, such as the annealed leap-point sampler, to conduct Bayesian inference for animal movement data and show this leads to new ecological discoveries as well.

A Unified Approach for Inferring Chemical Abundances from Stellar Spectroscopy and Images

Supervised by: Joshua Speagle

Stars are born in clusters that arise from dense clouds of molecular gas. While these stars initially stick together, over time these clusters end up dissolving and the individual member stars disperse throughout the Galaxy, mixing together with stars born in many other clusters. Each star, however, retains a unique chemical "fingerprint" based on the cluster they were born in. Combined with information on the properties of each star and its current path through the Galaxy, these chemical fingerprints (i.e. elemental abundances) allow astronomers to "wind the clock" back to uncover our Galaxy's evolutionary history.

There are a number of ongoing surveys that astronomers at UofT are involved in, each with their own particular approach (and often private code) for measuring abundances and other stellar properties from spectroscopy and imaging data. This project will involve a combination of (1) building a generalizable statistical framework for inferring these quantities that can be applied to many different datasets, (2) developing new codebases/improving existing (publicly available) ones, and/or (3) performing novel "cross-survey" code tests by working to adapt and apply a codebase developed for data collected by one survey (S5) to data collected by another (SDSS-V).

Predictive Bayesian Model Comparison in Astronomical Applications

Supervised by: Joshua Speagle

Many applications involve trying to compare various models and trying to find the "correct" one that truly characterizes the data or offers the most physical insight. Principles like "Occam's Razor" are widely taught as a general heuristic to address these issues, and hypothesis (model) testing itself is a huge part of many Frequentist inference approaches. However, within a Bayesian inference paradigm (where we have "beliefs" rather than "frequencies" of outcomes), performing model comparisons is often computationally and philosophically more challenging since we need often to marginalize over a wide range of prior beliefs (and many parameters!) when trying to draw a specific conclusion.

A set of recent approaches to address this issue has focused on developing and applying "predictive" criteria, i.e. choosing models based on how well they could potentially predict unseen data. This project would build on previous work to try and introduce these methods for use in astronomical applications. It would involve (1) becoming familiar with past work through a statistical literature review, (2) developing intuition, explanations, and visualizations through the use of simple analytical models that could be used to explain these concepts to astronomers, and/or (3) an application of this method to a real astronomical data set involving globular clusters.

Computationally Fast and Provable Subject (Non)Overlapping Clustering

Supervised by: Xin Bing

Clustering of subjects is an important unsupervised learning problem and has a central role in various applications of modern statistics. However, existing methodology either can only handle non-overlapping clustering task, that is, each subject is assigned to and only to one group, or does not enjoy theoretical guarantees for overlapping clustering, that is, each subject is allowed to belong to multiple groups.

In this project, we will develop a new clustering algorithm that has the following advantages: 1) it can handle both non-overlapping and overlapping clustering problems; 2) it has theoretical guarantees in terms of misclassification rate; 3) it is computationally fast and scalable for large data sets. The main idea of the new algorithm is to leverage a recently proposed score function, introduced in Bing, Bunea and Wegkamp (2022) for model-based latent factor models, to the context of subject clustering. Both different algorithm and theoretical analysis are expected.

Student Eligibility:

1. Must be an undergraduate student currently enrolled in a Department of Statistical Sciences Specialist or Major program.
2. Currently registered full-time or part-time student at the time of application.
3. Must have completed STA302 with at least a B+ grade.

How to submit your application:

Please fill out and submit the following application form:
Application for DoSS Summer Undergraduate Research Awards 2022

If you have any questions regarding these awards, please contact ug.statistics@utstat.utoronto.ca.

Completed applications are due by 11:59PM EST Sunday, March 19, 2022