URPS 2020: Undergraduate Research Program in Statistics for Winter 2020

Director of Undergraduate Programs: Prof. Edward Ionides

Undergraduate Program Coordinator: Gina Cornacchia

- Is this program for me?
- The application process.
- The projects.
- What to expect if I join a project.
- Do I get course credit? Can I use the project for an honors thesis? For the Data Science major capstone requirement?
- Other opportunities for undergraduate research in statistics.

Twitter and opinion polls

Relationships found between data extracted from social media and public opinion polls have led to optimism about supplementing traditional surveys with new sources of data. The goal of this project is to find and improve relationships between Twitter and public opinion polls using new topic modeling methods. The ultimate goal is to improve the accuracy of the public opinion polls (e.g., presidential approval, etc.), as well as "now-casting", in which public sentiment can be monitored in real time. Tweets can first be classified into various categories based on their content. Then the sentiment of those categories can then be related to existing relevant public opinion polls. Knowledge of Python is helpful.

Supervisors: Prof. Johann Gagnon-Bartsch and Robyn Ferg

Cancer drug screening

The student researcher will analyze large, complex datasets from two cancer drug screening experiments. The datasets will include information on the effectiveness of hundreds of drugs on hundreds of different cell lines, in addition to genomic information on the cell lines. The drug screening data contains widespread measurement error, which causes problems during analysis. With the ultimate goal of improving personalized cancer treatment, the student researcher will adapt and improve methods of measurement error detection and build prediction algorithms to determine which drugs are most effective against which types of cancers. The student researcher will learn to work with a variety of real-world, messy data (e.g. gene expression), methods to integrate different types of complex data, and various machine learning algorithms. Basic knowledge of R is required, and the student should expect to learn more R in the course of the project.

Supervisors: Prof. Johann Gagnon-Bartsch and Zoe Rehnberg

High-dimensional classification

This project will focus on exploring high-dimensional classification techniques. When the number of predictors exceeds the number of observations, traditional classification methods can perform poorly (or may not even work at all). We will be testing the effectiveness of a new high-dimensional classification method compared to existing methods. The undergraduate researcher will help to design simulation studies and work with various high-dimensional classification techniques. These methods will be applied to both the simulated data and high-dimensional genomics data. This subproject will require familiarity with R and/or Python as well as some familiarity with classification. As an alternate option, the undergraduate researcher may also choose to develop a faster implementation of the new classification method using the `Rcpp` package. This (optional) subproject will require the student to have familiarity with R and C++.

Supervisors: Prof. Johann Gagnon-Bartsch and Ed Wu

Optimal matching for impact evaluation: duality with & without certain matching restrictions

To estimate the benefit of a new curricular, medical or public policy intervention relative to usual practice, a common tactic is to begin by pairing individuals exposed to the intervention to otherwise similar controls. Pairing subjects well often requires use of constrained optimization routines. The algorithms in wide use – some of which are maintained by this research group – do not always scale well to large problems. This primarily mathematical project aims to help them scale up better. In this project, the student will begin with guided self-study of relevant network flow optimization methods. (If you've learned about Lagrange multipliers already, you're halfway there.) She or he will then go on to investigate and refine certain conjectures about what happens when after you've optimized against one set of constraints, you go back and revise certain of the constraints. There will be opportunities to use this work in impact estimation problems arising in education, medicine and/or public policy.

Supervisors: Dr. Mark Fredrickson and Prof. Ben Hansen

Optimal matching for impact evaluation: big data matching problems

To estimate the benefit of a new curricular, medical or public policy intervention relative to usual practice, a common tactic is to begin by pairing individuals exposed to the intervention to otherwise similar controls. Pairing subjects well often requires use of constrained optimization routines. The algorithms in wide use – some of which are maintained by this research group – do not always scale well to large problems. This computing and data analysis-oriented project would prototype updates to our R package, `optmatch`, that promise to improve its computational efficiency, going on to demonstrate the improvement in one or more replications of matching-based medical or social science impact analyses. A solid foundation in R and software development is required.

Supervisors: Dr. Mark Fredrickson and Prof. Ben Hansen

Predicting your likely post-secondary educational achievement on the basis of your high school

Our research group seeks qualified undergraduates to work on a project utilizing publicly available school-level data on Michigan high schools. The participant(s) will use multivariate regression models to construct indices aiming to capture students' likelihoods of enrolling in college, completing two or more years of college, and finishing college, among other post-secondary educational attainments. These indices will in turn play a supporting role to subsequent analyses evaluating curricula and programs within the University of Michigan. The current project will culminate in three interrelated related research products: 3-6 regression models to "predict" post-secondary attainment in Michigan as a function of high school characteristics; a reproducible program in R or Python that fits the model and regenerates key model diagnostics; and a report explaining your models and modeling decisions.

Supervisors: Prof. Ben Hansen and Tim Lycurgus

Opting out of M-STEP, Michigan's statewide K-12 achievement test

As statewide standardized testing has become more common, there has been a corresponding increase in "opting out" – parents or teachers simply declining to let their children be tested. The incomplete administrative data that result from this present new challenges for researchers evaluating education initiatives. To help meet these challenges, our research group is seeking qualified undergraduate students to help model the probability that a student will take the state test, as a function of the school a student attends and its aggregate student characteristics. The project will culminate in three interrelated related research products: 1-3 regression models to "predict" students' opting out of the M-STEP exam as a function of school-level characteristics; a reproducible program in R or Python that fits the model and regenerates key model diagnostics; and a report addressed to a lay audience describing how high-opt out schools tend to differ from low-opt out schools.

Supervisors: Prof. Ben Hansen and Tim Lycurgus

Modeling and data analysis to understand spatiotemporal epidemiology of dengue virus

Dengue fever is an emerging infectious disease which is now widespread through Africa, Asia, South America and Central America. Spatiotemporal patterns of dengue incidence are hard to explain using existing epidemiological models. This suggests there are gaps in our scientific understanding. New models will be proposed and computationally intensive statistical inference methods will be used to examine their success. The research project involves participating in a team of scientists and statisticians.

Supervisors: Prof. Edward Ionides and Kidus Asfaw

Bail bond companies and courthouses in the American legal system

This research project investigates the extent to which the presence of a courthouse attracts bail bond companies. These companies have a symbiotic relationship with the legal system; however, little is known about their geographic clustering behavior around local courthouses. As part of this research, we need to build a database of existing bail bond companies, courthouses, and the areas they serve. The undergraduate researcher will assist with implementing the software for gathering this data from the web and subsequent data preparation and cleaning.

Supervisor: Dr. Keith Levin

Neuroscience of addiction

The data are from a neuroscience pilot experiment in the lab of Prof. Shelly Flagel (Psychiatry), investigating the links between brain chemistry and addiction in rats. The aim of the project is to predict, based on levels of different chemicals in the brain, whether or not a given rat will display different types of behaviors associated with addiction after several days of conditioning experiments. The student will carry out data analysis building on methods learned in STATS 413 and 415.

Supervisors: Prof. Liza Levina and Dr. Keith Levin

Human Brain Connectivity

This project is motivated by problems in neuroscience where data from the brain are represented as both networks of connectivity patterns (pairwise connections between multiple locations in the brain) and measurements of other covariates at these same locations in the brain. We have developed a predictive method that exploits structure in the data and can offer greater scientific interpretability relative to standard approaches. The undergraduate researcher will help apply this method to multiple human brain imaging datasets from neuroscience collaborators. You will be predicting individual characteristics (like psychiatric diagnosis or IQ) using functional connectivity and structural brain measures obtained using magnetic resonance imaging (MRI). The student should have at least one of STATS 413 or STATS 415 (prior or concurrent) and be proficient in R. An interest in and some knowledge of neuroscience is a plus but not required.

Supervisors: Prof. Liza Levina and Dan Kessler

Reinforcement Learning

We will study the book "Reinforcement Learning" by Barto and Sutton, in a reading group focusing on bandit problems. The classic problem in the field is as follows: suppose you have k slot machines and only n rounds to play but you don't know the probability of success on each machine, how do you optimize your expected winnings?

Supervisor: Dr. Asad Lodhia

Hot hands: Shooting streaks in basketball

This project concerns the hot hand phenomenon in basketball, the belief that players have streaks in shooting (Gilovich,Vallone & Tversky(1985)). Miller & Sanjurjo (2018) recently showed that the analysis of Gilovich et al. (1985) introduces non-negligible (statistical) bias in estimating the probability of success given that the previous outcome was a success. By analyzing the data of Gilovich and utilizing computer simulations, we will investigate the bias and consider alternative approaches to estimating this conditional probability.

Gilovich, T., Vallone, R., & Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. Cognitive psychology, 17(3), 295–314.

Miller, J. B. & Sanjurjo, A. (2018). Surprised by the hot hand fallacy? A truth in the law of small numbers. Econometrica, 86(6), 2019–2047.

Supervisors: Prof. Ya'acov Ritov and Michael Law

Analysis and Visualization of Darknet Internet Traffic

The project involves designing and developing an interactive web-application via the R Shiny platform for the analysis of Darknet internet traffic. The Darknet is traffic routed to the space of unassigned (dark) part of the network, not to be confused with the "Darkweb", a collection of sites for illicit activity. Darknet traffic originates from 1) misconfigured or malicious hosts who are scanning for cybersecurity vulnerabilities in the network, 2) from randomly spoofed IP packets aiming to attack specific victims (the so-termed 'backscatter' traffic) and 3) from networking misconfigurations. The dashboard will access a Big Query Google data-base, which contains real-time Darknet traffic measurements. It will compute real-time summary statistics and cybersecurity threat indices, which quantify, classify, and potentially localize cybersecurity threats to the network. The project will train the student in statistical methods such as exploratory data analysis, principal component analysis, extreme value theory, clustering, and classification.

Supervisors: Prof. Stilian Stoev and Dr. Michalis Kallitsis (Merit Network)

Active learning in the streaming setting with purely random trees

Imagine you are in a situation where acquiring unlabelled data is cheap, but labelling them is expensive; for example scraping images, text or audio from the internet is cheap, but requiring humans to label them is expensive. As a result you can only request a small amount of the data to be labelled, and you want to build the best statistical model you can with that finite budget of labels. How do you iteratively select the best new data point i to label, using the unlabelled data you start with, plus the labels of data points $1, 2, \ldots (i-1)$? This is the goal in active learning: to develop algorithms which automatically decide what data should be labelled.

Supervisors: Prof. Ambuj Tewari and Jonathan Goetz