# Online Proceedings

## POSTER SESSION 1

**MGH 241, Easel 130**

*11:00 AM to 1:00 PM*

**Developing a Shiny App for Investigating Fatal Encounters with Police**

*Madeline (Maddi) Cummins, Junior, Informatics*
*Marwa (marwa) Elatrache, Junior, Exchange - Arts & Sciences*
*Mentor: Martina Morris, Statistics and Sociology*

National interest in the fatal shooting of civilians by police is growing, driven in part by high profile cases that are now often captured on video. Several news organizations have done in-depth reporting on local or national trends. While the data these organizations used has been made public, the technical skills needed to access and analyze the data present a barrier to public use. Last year we had a prototype for an app to increase access and the ability for anyone to investigate fatal encounters with police, this year we have a finalized package. This package includes a browser-based software application to support access to the data, along with tools for Exploratory Data Analysis (EDA) to support exploration of the data. For this project we worked in the programming language R, writing a shiny app to provide the browser-based interface to R's powerful EDA tools. Shiny apps require two software components: underlying R code for analyzing the data, and a user interface (UI) that provides simple point and click tools for manipulating and running the code. For the R code, we focused on graphical exploration of the data: time series charts at different levels of spatial aggregation, interactive maps where users can zoom in and view online links for individual cases, and animated cartograms. For the UI, we developed an app with tabs for each graphical option, and a range of filters and aggregation options on each tab. Users can easily view and interact with each visual and modify it to display what they are interested in. Our codebase follows the guidelines for reproducible research and is uploaded to a public GitHub repository. This supports both public use and public development; users interested in contributing to the codebase can clone our repository, and submit suggested edits via pull requests.

## POSTER SESSION 1

**MGH 241, Easel 133**

*11:00 AM to 1:00 PM*

**Where's the Last Fish? A Bayesian Hierarchical Model to Characterize Sampling Effort in Single-Pass Electrofishing Surveys**

*Emily Flanagan, Senior, Statistics, Mathematics (Comprehensive)*
*Connie He, Senior, Statistics*
*Mentor: Tamre Cardoso, Statistics*

Washington State Forest Practice Rules require wider riparian buffer zones adjacent to streams that support fish versus those that do not. Consequently, verification of fish presence in streams is an important pre-timber harvest requirement of landowners. These determinations are made by professional biologists using single pass electrofishing protocols prescribed in state policy. The protocol stipulates that fish are presumed absent if no fish are found after sampling 1,320 feet beyond the last detected fish. What confidence should be placed in this level of effort? To better characterize the confidence in last fish detections, we implement a Bayesian hierarchical model, using distance to first fish as a measure of sampling effort, to estimate the distribution required to find fish using single pass electrofishing methods. We use the model output to determine the distribution of 99th percentile distances for first fish detections given fish are present in the stream. Further, the distribution of 99th percentile distances can be used to establish confidence in the boundary between fish presence and absence for single pass electrofishing surveys. Other percentile distances could be selected by policy makers to describe desired confidence levels. The model was developed using distances to first fish detections collected on 108 randomly selected streams from western Washington. Ancillary data included the length of the first detected fish, wetted and active channel widths, and stream gradients.

# SESSION 1L

## MATHEMATICAL MODELING IN THE SCIENCES

*Session Moderator: Elizabeth Thompson, Statistics*
**MGH 271**
*12:30 PM to 2:15 PM*
* Note: Titles in order of presentation.

**Finding Clusters in Data Through Recursive Graph Thresholding**
*Huanbiao (Richard) Zhu, Sophomore, Statistics*
*Mentor: Werner Stuetzle, Statistics*

The goal of clustering is to partition a dataset into subsets ("clusters") such that observations in the same cluster are similar and dissimilar from observations in other clusters. We think of the clustering problem as a graph problem. The vertices of the graph are the observations. Any two vertices are connected by an edge. The weight of the edge is the dissimilarity between the observations. We can form clusters by progressive graph thresholding. We repeatedly break the longest edge until the graph has two connected components. These two connected components form clusters in the sense described above. We then apply the same procedure recursively until we have obtained the desired number of clusters. We propose a method to efficiently find the connected components and their memberships.

# SESSION 1L

## MATHEMATICAL MODELING IN THE SCIENCES

*Session Moderator: Elizabeth Thompson, Statistics*
**MGH 271**
*12:30 PM to 2:15 PM*
* Note: Titles in order of presentation.

**How do Mean and Variance Affect Gene Survival and Gene Frequencies?**
*Jueyi Liu, Senior, Economics, Applied & Computational Mathematical Sciences (Scientific Computing & Numerical Algorithms)*
*UW Honors Program*
*Mentor: Elizabeth Thompson, Statistics*

Mutation produces new variation in populations, and in each generation these variants are copied from parents to offspring. While almost all variants of genes are lost, they may remain in the population for many generations. We use branching process models to analyze counts of gene copies. In a population of constant size, on average, a gene copy produces one offspring copy at the next generation. An advantageous mutant will have a mean greater than 1, and a deleterious one will have a mean less than 1. It is thought that most mutations are slightly deleterious, and with high probability those variants become extinct rapidly. Nonetheless, the few deleterious mutants that are not yet extinct may achieve high numbers. Thus, we have a particular interest in those with a mean slightly less than 1. We use different probability models for the offspring distribution and consider the mutant's survival about: the extinction probability over $k$ generations, the expected copy count conditional on survival, and the probability of survival additional $k$ generations conditional on surviving $k$ already. We find that variances in addition to means of offspring distributions closely relate these statistics. By adjusting parameters of distributions, we let the mean and variance be approximately the same across distributions. Based on our simulations, when $k$ is large and the mean and variance of offspring are the same, the mutant's survival condition is uniform throughout. In other words, those statistics above can be estimated by the mean and variance exclusively, and the specific distribution does not affect much the conditional population dynamics. However, at the first few generations, these statistics are different for each distribution. Thus, if we know the mean and variance of a mutant, we can predict the long-term population behaviors conditional on survival without knowing the true distribution of the mutant.

# SESSION 2S

## THE POWER OF MEDIA REPRESENTATIONS AND DIGITAL ARCHIVES

*Session Moderator: Carmen Gonzalez, Communication*
**JHN 175**
*3:30 PM to 5:15 PM*
* Note: Titles in order of presentation.

**Record Linkage and Multiple Imputation across Geospatial and Demographic Data on Fatal Police Encounters**
*Vaughn Isaac Johnson, Senior, Statistics*
*Mentor: Martina Morris, Statistics and Sociology*

Every year, hundreds of people die from fatal police violence, and there is no official repository that records this information. Over the past decade several crowd-sourced efforts have emerged to fill this gap, creating online repositories that compile the geospatial and demographic information of victims of fatal police force. Our team is working with three of these. The oldest and largest data set has information dating back to 2000, and contains roughly 17,000 observations, while the

other two date back to 2013, and closer to 4,400 observations. There is missing data in all of the repositories, and the missingness levels are particularly high for the race of the victim, a variable of interest. We know a priori that there is substantial overlap in the cases covered by these three data sets, so we hope to use record linkage methods to combine the information across the data sets to recover or impute the missing data. Previously, we harmonized the three data sets, so the problem of inconsistencies in variable names, formatting or other irregularities has been addressed. We now have four sequential, dependent goals. Our first goal is to perform record linkage within each data set to eliminate duplicate records. Our second goal is to perform record linkage across the three data sets, using the victim's state as a blocking key to reduce the computational burden. Our third goal is to address the remaining missing values through statistical imputation. Our final goal is to provide a public repository with a clean unified data set, and the code needed to reproduce this from the original raw data. This will be paired with a corresponding browser-based public tool for data exploration in the form of an online R Shiny App.

## POSTER SESSION 4

**MGH 241, Easel 144**

*4:00 PM to 6:00 PM*

**Evaluating the Performance of Sea Ice Models After Bias Correction**

*Zihui (Joyce) Zhang, Senior, Applied & Computational Mathematical Sciences (Discrete Mathematics & Algorithms), Statistics*
*Mentor: Adrian Raftery, Statistics*
*Mentor: Hannah Director, Statistics*

Sea ice, frozen ocean water, in the Arctic is declining due to climate change. This is getting more attention, since sea ice can substantially affect wildlife, ecosystems, and human society. Sea ice extent, a measure of the surface area of the ocean covered by sea ice, is used to monitor the environment. A good sea ice extent model can benefit not only studies of the Arctic but also the global economy, as many ships use the Arctic area as part of their route. There are various sea ice extent models in the science community that give good estimates of sea ice extent. However, there are instances where their predictions do not match observations. My research explores bias correction methods that help predictions to match observations and compares the performance of sea ice models after bias correction. In this analysis, I consider the performance of a sea ice model produced by the Geophysical Fluid Dynamics Laboratory using boxplots and rank histograms. This work helps show the value of bias correction. I will compare the performence of the models to identify in what month and in what location models predict well. This information will support improved prediction of the Arctic going forward.

## POSTER SESSION 4

**MGH 241, Easel 145**

*4:00 PM to 6:00 PM*

**A Comparison of Variability Due to Cross-Validation and Initialization in Neural Networks**

*Jueyi Liu, Senior, Economics, Applied & Computational Mathematical Sciences (Scientific Computing & Numerical Algorithms)*
*UW Honors Program*
*Mentor: Caren Marzban, Statistics, and Applied Physics Lab*

It is well known that nonlinear optimization can lead to a local minimum of the loss function. As such, different initial values of the model parameters can give different values for the loss function. In other words, the existence of local minima introduces a source of variability in the loss function. Additionally, model selection often involves resampling, which in turn introduces a second source of variability. In this work, random effects models are employed to estimate these two variance components. More specifically, a neural network model is employed to examine the behavior of these variance components as a function of the variance of the initial weights and the number of hidden nodes (H). It is found that when H is small, weight initialization has a larger effect on variation of loss than cross-validation, and when H is large, these two factors contribute comparably to the variability in loss.