Online Proceedings

SESSION 1B

DATA SCIENCE, STATISTICS AND SOCIETY

Session Moderator: , Statistics and Sociology
MGH 082A
12:30 PM to 2:15 PM
* Note: Titles in order of presentation.

Fatal Encounters with Police: Improving Public Access to Exploratory Data Analytics
Madeline Ann (Maddi) Cummins, Sophomore, Pre Engineering
Mentor: Martina Morris, Statistics and Sociology
Mentor: Ben Marwick, Anthropology

National interest in the fatal shooting of civilians by police is growing, driven in part by several high profile cases that were captured on video. Several news organizations, including the Washington Post and the Seattle Times, 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. This project seeks to develop browser-based software that will support simple access to the data, along with tools for Exploratory Data Analysis (EDA), that will make it easier for others to learn from these data. For this project we are working in the programming language R, writing a shiny app to provide the browser-based interface to R's powerful EDA tools. Shiny apps involve writing two software components: the underlying R code for analyzing the data, and a user interface (UI) that provides a simple point and click tools for running the code. For the R code, we will focus 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, choropleth maps to visualize racial disparities, and animated cartograms. For the UI, we will develop a webpage with tabs for each graphical option, and a range of filters and aggregation options on each tab. Users will be able to easily view and interact with each visual and modify it to display what they are interested in. Our codebase will follow the guidelines for reproducible research, and be uploaded to a public GitHub repository. This will support both public use and public development; users interested in contributing to the codebase can clone our repository, and submit suggested edits via pull requests.

SESSION 1B

DATA SCIENCE, STATISTICS AND SOCIETY

Session Moderator: , Statistics and Sociology
MGH 082A
12:30 PM to 2:15 PM
* Note: Titles in order of presentation.

Matching Heterogenous Datasets Using Seeded Classification
Jainul Vaghasia, Sophomore, Statistics, Computer Science
Mentor: Martina Morris, Statistics and Sociology
Mentor: Ben Marwick, Anthropology

For those interested in understanding more about fatal shootings of civilians by police, a natural place to start would be data on the events. Such data is, however, surprisingly hard to access. There are no official sources, but there are several crowd sourced data sets available online. These data sets each contain some common and some unique fields, and they are stored in very different formats. Hence, these data sets require preprocessing before the data can be used for other practical purposes like analysis and visualization. In this research, we focus on cleaning such heterogenous data sets using statistical methods with major focus on data matching. Data matching refers to matching data from different sources and recognizing the matching entities. It is an essential step in data cleaning that helps in resolving the conflicting information provided by the data sets—clerical errors, missing data, differently stored data, etc. Traditional statistical procedures and classification algorithms are based on supervised learning which requires training data. However, it is quite common that the training data is not available at hand and it might not be cost effective to prepare one from scratch. We explore a two step unsupervised learning algorithm that might achieve the same performance as the supervised counterparts. In the first step, it prepares a seed data set containing data points that are relatively extreme matches and non-matches based on discriminating fields. This prepared data set can be used as training data to train classical classification algorithms—k-nearest neighbors, Support Vec-

tor machines, Random Forests—which in turn can be used to classify the remaining data points. In the process, we examine missing data, geocode matching, and feature selection methods that increase classification accuracy by selecting the most discriminating fields. Finally, we examine the prospect of generalizing this approach to suit differently aligned data.

POSTER SESSION 3

Commons West, Easel 40

2:30 PM to 4:00 PM

More and Larger Snowstorms: Big Fake News

Yi Alan (Alan) Zhang, Senior, Mathematics, Economics
Xinyue Li, Senior, Economics, Statistics
Yuqun (Azura) Tang, Senior, Economics, Statistics
Mentor: Peter Guttorp, Statistics

Recently, there seems to be news about blizzards striking East Coast cities nearly every year. However, most reports simply claim that the frequency and intensity of large snowstorms on the East Coast are increasing without providing convincing evidence. Are they telling the truth? We decided to statistically test if the frequency and intensity of large snowstorms in coastal North America have increased in recent years. In our research, we encountered many difficulties during the data collection and cleaning processes. For example, most cities only have records since around 1950 and the Canadian snow data are hard to clean. We first plot frequency of large snowstorms for each city and detected that there is actually a downward trend of snowstorm frequency in coastal North America. To further verify our finding, we also permuted the time-series data and performed permutation tests on the slopes of the linear regression of extreme snowstorm frequencies versus year. In our current progress, the permutation tests suggest that, for all East Coast cities studied, the hypothesized increases in annual frequencies of extreme snowfall are not statistically significant at any reasonable level of significance for looking at exceeding both the 75% and 90% quantiles of annual snowfall. Thus, our current results imply that the frequency of large snowstorms in many East Coast cities has not been increasing in recent years, which is different from the media claims.