Data and Methods

Adam Shelton

12/6/2019

Data

This project primarily utilizes data from two sources, the City of Chicago and The University of Chicago. Crime data from the City of Chicago is available through the city's public data portal. This data is extracted from the Chicago Police Department's Citizen Law Enforcement Analysis and Reporting system (City of Chicago 2011). The data is based of initial reports of crimes, as gathered by the Chicago Police Department, and does not appear to be updated or revised if any information on a reported crime changes later. This also means that information about crimes, if reported to CPD incorrectly, would likely be recorded correctly in this data-set. From the city's data portal, this data-set can be directly downloaded as a CSV or queried through an API.

The University of Chicago publicly publishes three data-sets about the University of Chicago Police Department on the university's Department of Safety and Security website. The most analogous of these data-sets to the data on reported crimes published by CPD are UCPD's incident reports. Unlike the crime data provided by the CPD, this data-set also includes other reports of emergencies, such as fires and medical emergencies, that the UCPD responded to. The University of Chicago also publishes more granular information on the specific stops of vehicles and people, which they call traffic stops and field interviews, respectively.

While the University of Chicago does publish this data publicly, it does not come in an easily aggregated form, like crime reports from the city do. The data is available on the university's Department of Safety and Security website, with each of the three available data-sets on a different page. This interactive webpage allows for the user to specify a date range of archived data, and then displays five observations on a page for the user to navigate through. As this data is displayed in a table on the webpage, it is a rather trivial task to web-scrape the data using tools such as `rvest` for R. The web archive uses URL queries to specify a date range and a page of observations, and after a little trial and error to find the date of the first reported data, it is simple to query the archive for any page in the entire archive. Then the data must be scraped and appended to a data-frame, one page of five observations at a time from the table, until a full data-set has been obtained.

Crime data from the City of Chicago has been used extensively as one of the first data-sets released on the open data portal, with over 500 thousand downloads since its inception (Goldstein and Dyson 2013; City of Chicago 2011). Crime in Chicago and its causes and effects have been studied quite extensively as well, with data from the city's data portal being used in research by PhD students and full research organizations alike, which has even informed policing policy in the city (City of Chicago 2011). However, much less attention has been given to the data published by the University of Chicago. While some analysis of stops by UCPD has been done, it appears very little research has studied crime reports by the UCPD (Newman 2016).

Crime data from the City of Chicago encompasses approximately seven million reported crimes from 2001 to 7 days from the present for the entire city (City of Chicago 2011). It includes categorizations of crimes, and the time and location a reported crime occurred (City of Chicago 2011). As this study focuses on the interaction of CPD and UCPD this data would likely be subsetted to include a similar area as reflected in data from the University of Chicago.

Data on reported incidents is primarily contained to the defined "jurisdiction" of UCPD, which encompasses Hyde Park and five surrounding neighborhoods (Sherman 2019). However, while the university defines this jurisdiction as the area UCPD officers actively patrol, UCPD officers have the authority to operate anywhere in Cook County (Sherman 2019). This, coupled with the University of Chicago operating buildings much farther outside of the Hyde Park campus, results into the UCPD responding to or recording incidents that happen outside of their typical patrol area. The University began publishing data from the UCPD on a near daily basis starting on July 1, 2010 for incident reports and July 1, 2015 for traffic stops and field interviews. In this time period, there have been approximately 11,000 incident reports, 4,000 traffic stops, and 1,500 field interviews.

Table 1: Overview of Data-sets

Institution	Data Type	Columns	Rows (approx.)
City of Chicago	Crime reports	22	7 million
UChicago	Incident reports	7	11,000
UChicago	Traffic stops	9	4,000
UChicago	Field Interviews	8	1,500

Additional sources of data may be gathered to supplement the analysis, such as demographic data from the US Census API, national data-sets on crime from federal organisations like the Federal Bureau of Investigation, and non-public information like the full text of police reports from the City of Chicago through a Freedom of Information Act request or contact with the Office of the Inspector General. Additional data would primarily be used to gather data on the racial backgrounds of officers and perpetrators of crimes. Incident reports from UCPD and CPD do not formally contain information about race, age, or even gender, and would need to be inferred by the location of the incident, or gathered from a different data-source.

Methods

Data mining techniques such as unsupervised machine learning models are used alongside an exploratory data analysis to provide more insight into the reports that each police department handles within the jurisdiction of the UCPD. In large data-sets, it is not always immediately clear what relationships or groupings are in the data, especially with complicated issues such as crime and policing. Unsupervised clustering is used to reduce the feature space of the data from each respective department, to clarify what types of crimes are happening in which locations, or to discern whether a combined data-set of all reports from both departments can be reliably separated on their attributes alone.

The first part of the clustering process involves extensive visual analyses of the data to create a breakdown of reports in the area and ascertain clusterability and assist in selecting relevant features. Many clustering methods such as K-means, AGNES, PAM, and DBSAN would be used and assessed to arrive at the best performing model. Due to significant number of categorical variables in the data, it would be most appropriate to use Gower's distance instead of Euclidean distance in these models.

As UCPD incident reports also include narrative data in the form of short descriptions of the reported incident and UCPD's response to the incident, text mining methods are used to better understand structural forces that may affect how the UCPD reacts to incidents. Topic modeling using Latent Dirichlet allocation or Doc2Vec works to explain what is generally discussed in incident reports, and how certain incidents, even those from the same category, may be different in ways not captured with other variables. Results from any text mining will be used to engineer features that an unsupervised modeling method could utilize.

With the addition of other data, such as engineered features from any text data, or demographic data from the US Census, it also becomes increasingly important that only the most relevant measures are used in any unsupervised clustering models. Adding too many variables to a model be it supervised or unsupervised can hurt its performance and accuracy. This necessitates rigorous dimension techniques like Principal Components Analysis or factor analysis, especially when considering the many demographic variables available through the census. PCA or factor analysis first are used to assess whether the relationships in the data line up with expectations based on previous research. This provided insight on areas where the data may not be complete enough, or interesting relationships that had not been consider before that warrant additional research.

The results of these analyses not only gives evidence on whether variables of interest should be included or excluded, but also more opportunity to improve the features used in an unsupervised clustering model. It is expected that many variables might measure latent features that we do not have a direct measure for, like socioeconomic status or inequality, but we can get much closer by combining the right variables in the right proportions, which confirmatory factor analysis handles quite well.

A variety of modeling methods are then used to predict a series of dependent variables to more explicitly understand the differences between both departments. This will provide more evidence for how each department uniquely responds to reports of crimes and what possible ramifications for police and citizens this could impose. A dependent variable specifying the department of origin for a specified report is the first step. Other models are also built to predict the type of crime, and whether an arrest was made, to further explore any possible differences between departments. Independent variables for these models would include the time and location of crimes, racial and demographic measures of the area where the crime occurred, including proportions of gender, race, residents renting/owning, residents with a house loan, children, families, vacant residences, median age, and population density

The modeling algorithms used have to be able to predict categorical outcomes, which rules out typical linear regression. Therefore, a logistic model could be used to model only a binary outcome between either department. More complicated methods such as Linear or Quadratic Discriminant Analysis, Naive Bayes, K-Nearest Neighbors, Decision Trees, and Random Forests are used and their performance compared to pick the best models to analyze. In the case of less interpretable models, such as Random Forests, partial dependence plots and individual conditional expectation plots are used to better interpret and assess the legitimacy of the relationships displayed by the model.

References

City of Chicago. 2011. “Crimes - 2001 to present.” https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2.

Goldstein, Brett, and Lauren Dyson. 2013. *Beyond Transparency: Open Data and the Future of Civic Innocation*. https://doi.org/10.1017/CBO9781107415324.004.

Newman, Jonah. 2016. “New data supports old accusations of racial profiling by University of Chicago Police Department.” https://www.chicagoreporter.com/new-data-supports-old-accusations-of-racial-profiling-by-university-of-chica

Sherman, Stephen Averill. 2019. “From Revanchism to Inclusion: Institutional Forms of Planning and Police in Hyde Park, Chicago.” *Journal of Planning Education and Research*, 0739456X1987768. https://doi.org/10.1177/0739456x19877683.