CS 229 - Final Project Proposal

Kane, Alyson
alykane@stanford.edu
SUID #06079278

Sagalovsky, Ariel
asagalov@stanford.edu
SUID #05519028

October 21, 2016

1 Topic Idea

The main purpose of our project is to gain insights into crime patterns across major U.S. cities. We are interested in understanding which demographic factors may contribute to neighborhood level crime, and if these factors are consistent across cities. We will implement both supervised and unsupervised learning techniques to build predictive models and identify key factors that contribute to crime rates.

2 Data Sources

We will be compiling a handful of data sources to build our training and test sets. For each city, we will download logs of reported incidents of crime through the OpenData portal, which include the time and date of occurrence, the type of crime, and a handful of geographic identifiers (block, neighborhood, latitude, longitude)[1].

In addition, we will augment the crime data with the most recent Census data at the Block Group level to identify demographic features (ethnicity, income, educational attainment, poverty levels, etc.) for each neighborhood in our study[2]. We will pull crime and demographic statistics across major US cities including but not limited to New York, Los Angeles, Chicago, San Francisco, Philadelphia, and Washington, D.C.

3 Methodology

We seek to achieve our goal using three approaches:

1. Cluster neighborhoods across different cities by demographic attributes
2. Time series modeling by individual neighborhoods
3. Modeling each US city in our analysis individually to understand distributions of crime reports by geography

After building models using these distinct methods, we wish to expand on our research by extracting relevant features using variable importance calculations and independent component analysis.

[1] https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2/data

[2] http://factfinder.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t

For instance, we may wish to know if in using 2010 Census data for demographic breakdowns, are our models more accurate for predicting crimes in, say 2011, than 2015? If so, we may be able to single out specific neighborhoods for changes in overall crime rates. In doing so, we can identify specific geographies that have undergone gentrification or decline. Further, we can glean which inputs to our model have decreasing predictive power over time. These features can serve in a subsequent analysis of how to predict which areas are undergoing the most change.

We may also seek to identify how demographic factors, such as ethnic segregation, in a city may lead to the localization of crimes. For instance, the map below depicts how residents of particular ethnicities are distributed in New York and Chicago[3]. Comparing the maps below to crime data, will we see differences in crime types or frequencies across similar segmentations?

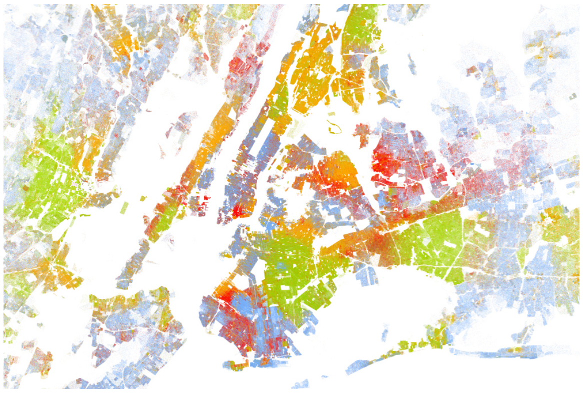

Figure 1: New York

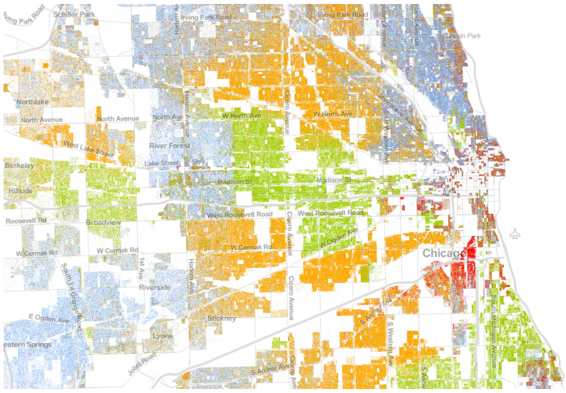

Figure 2: Chicago

[3] https://www.wired.com/2013/08/how-segregated-is-your-city-this-eye-opening-map-shows-you/