Spatial Analysis of NYPD SQF Clustering in 2011

Alan Chen, Clare Clingain, Emily Coco, Wenshu Yang, Haowen Zheng

October 17, 2018

Introduction

In this project we wanted to determine whether or not the reasons given for stops in New York City indicate racial biases on the part of the officers. To address these questions, this project looks at the rate of stops because of clothing and because of furtive movements within census tracts. Therefore, we pose the following questions:

1. Where are the "hotspots" (by census tract) of people being stopped for i. clothing, ii. furtive movements?
2. What is the demographic breakdown of the census tracts these "hot spots" are happening in?
3. What is the relative probability of being stopped for i. clothing, ii. furtive movements within their respective clusters?

Materials and Methods

Data and Pre-processing

First, we acquired 2010 Census Tract polygons from NYC Open Data and demographic data from the American Community Survey. Race demographics were aggregated to match the Stop and Frisk racial categories. Percentages of each racial category in the total population were calculated. Census tract codes were standardized across files, which originally had different numbers for boroughs versus counties. The finalized racial demographic data was merged with the shape polygons.

Next, the times variable in the Stop and Frisk data frame was first converted to class chron from the chron package in order to subset the data between 10:00 p.m to 6:00 a.m. The dates of stops were converted to date objects, and subsequently subsetted to Sunday through Thursday.

The Census Tract and Stop and Frisk data were merged into one data frame. Any cases with missing latitude and longitude, or that occured in census tracts where no one lives and no stops were made (n = 13) were dropped. Variable names were updated for clarity. The final cleaned data set was saved as an .RData file.

Rates of stops for clothing and furtive movements were calculated for each census tract.

$$Rate_i = \sum_{i=1} \frac{stopped.x_i}{total.stops_i}$$

where i signifies each census tract. Thus, there were 2,155 rates for each variable of interest.