Spatiotemporal Analysis of Racial Bias in NYPD Stop, Question, and Frisk Procedures

Mihail Kaburis[1] | Samantha Kamath[2] | Jasmine Mdlock[3]
Advisor: Dr. Xinyue Ye[4] | Mentor: Jun Yuan (PhD Student)[4]
Department of Informatics
[1]University of South Florida, Tampa, FL 33620
[2]University of Miami, Coral Gables, FL 33146
[3]University of Maryland, Baltimore County, Baltimore MD 21250
[4]New Jersey Institute of Technology, Newark, NJ 07102

Abstract

The treatment of racial minorities by law enforcement personnel has become heavily debated in the United States. Determining the levels of racial bias during routine police stops could have tremendous implications on police-civilian relations. Our research examines racial bias within the NYPD's Stop, Question, and Frisk (SQF) procedures, which have disproportionately stopped black and hispanic individuals over their white counterparts. We determine racial bias via the threshold test, a Bayesian statistical model that calculates a race and precinct-specific search threshold — the point at which a police officer decides to search an individual for contraband. Previous threshold test models have not accounted for spatiotemporal factors, including the fluctuation of racial bias within police precincts over time and the amount of crime in the neighborhoods where stops occur. The NYPD's SQF Datasets and aggregated Historical NYC Crime Data from 2003 to 2018 were used to determine how racial bias in SQF procedures varies according to these factors. While there was no significant correlation between the search thresholds and neighborhood crime numbers, white search thresholds were consistently higher than black and Hispanic search thresholds, implying consistent racial bias against these racial minorities. This information, along with a robust visualization platform for racial bias, provides valuable tools for policymakers, police, and residents to make informed decisions about policing practices.

Introduction

Recent years have caused law enforcement interactions with racial minorities to come under heavy scrutiny by politicians, the press, and the general public. Racially-charged incidents involving police, such as derogatory Facebook posts written by officers in Philadelphia and the fatal shootings of black civilians across the U.S., have sparked controversy over the unfair treatment of minorities by law enforcement. These incidents have caused many to raise concerns about police officers unconsciously adopting racial bias when dealing with minority individuals. While implicit bias, the unconscious attribution of particular qualities of a certain social group, remains embedded in institutions related to law enforcement in the United States, policy makers and chiefs of police have increasingly focused on what is often called implicit bias, inherently unintentional yet more pervasive (GreenWald, 1995; Baker, 2018). If police officers rely on stereotypes rather than concrete facts, that bias can escalate the situation and have serious, and potentially lethal, consequences. In particular, the New York City Police Department's (NYPD) Stop, Question, and Frisk (SQF) procedures, in which police officers detain, question, and inspect individuals for contraband under reasonable suspicion, have been accused of racial discrimination.

The United States Supreme Court made an important ruling on stop and frisk practices in the 1968 case Terry v. Ohio. Until 1968, a police officer could search only someone who had been arrested, unless a search warrant had been obtained. However, the ruling granted permission for officers lacking probable cause for an arrest to frisk individuals if the officer has reasonable suspicion they may be dangerous (Katz, 2004). Since this ruling and subsequent court cases on unreasonable searches, the NYPD's implementation of stop and frisk procedures has caused a major debate around the country. As people became aware of bias within law enforcement, many have questioned the effectiveness of SQF, believing this method of policing targets young black men and increases crime rates. The NYPD's SQF procedure was found to disproportionately stop

blacks and Hispanics; when stopped, these racial groups were less likely than whites to possess a weapon, which implies that police officers apply racially-biased decision-making (Goel et. al., 2016). While New York City has launched police reform initiatives under the leadership of May Bill de Blasio, including diversity training programs for officers, improvements in policing practices in the city remain unknown (Baker, 2018).

Several different factors determine whether a police officer is being discriminatory, including race, behavioral signals, location, time, etc. In response, various bias tests have been created to assess this behavior. The model we utilize in this paper, the threshold test model, best considers the complexity of the New York City area and the SQF procedure. However, previous bias test models have not accounted for spatiotemporal factors, including the fluctuation of racial bias within police precincts over time and the amount of crime in the neighborhoods where stops occur. To better understand racially-discriminative police behavior and propose improved discrimination tests, we employ the threshold test model on NYPD SQF data from 2003 to 2018 for the purposes of determining how criteria for racially-discriminative SQF procedures varies depending on space and time.

Background

The Threshold Test

The threshold test model was originally developed in 2017 by researchers at Stanford University. Previous models, including the benchmark and outcome tests, did not consider the full spectrum of factors determining whether an incident was authentically biased. The benchmark test only compared the search rates of different racial groups, causing the model to have major limitations. For example, consider a region which is mostly populated by black and Hispanic racial groups. There exists a high probability that minorities will be stopped more often than white individuals. This is not necessarily indicative of racially biased policing practices. Addressing this shortcoming of the benchmark test, researchers proposed the outcome test, which is based on the success rate of searches rather than the total search rate. Researchers argued that even if different groups vary in their propensity to carry contraband, discrimination could be detected if searches of minorities yield contraband less often than searches of whites (Simoiu, 2018). However, this test failed to consider the possibility that different racial groups may have different risk distributions of carrying contraband; in this case, discrepancies in the hit rates of different racial groups may reflect this variance in risk distribution rather than indicating racially-biased policing.

The threshold test, however, combines information on both the search and hit rate and allows us to directly infer the standard of evidence officers require before carrying out a search. The test uses a Bayesian model to determine a race- and precinct-specific search threshold — a specific likelihood of carrying contraband at which an officer decides to search an individual. If an officer determines an individual's probability of carrying contraband is greater than this threshold, the individual is searched.

To visualize this Figure 1 graphs the hypothetical risk distributions (solid curves) and search thresholds (dashed vertical lines) of two different racial groups (represented by different colors). The search rate for a given group is equal to the area under the risk distribution but above the threshold, and the hit rate is the mean of the distribution conditional on being above the threshold (Simoiu, 2017).

These thresholds characterize the general standard of suspicion that a police officer utilizes to investigate an individual further. Search thresholds that are lower for certain racial groups are indicative of biased policing practices against these groups. Precincts can have varying thresholds for different races, and by comparing them, bias can be detected between racial groups.

The threshold test model utilizes Markov Chain Monte Carlo (MCMC) algorithms to take into account prior distributions and create a sample space indicative of thresholds for police departments and precincts. Specifically, for each department and race group, the model compares the observed search and hit rates to their expected values under the sample data-generating process with parameters that are drawn from the inferred posterior distribution. The sampling procedure yields 2,500 'warmup' draws from the joint posterior distribution of the parameters. For each parameter draw, the model analytically computes the resulting search and hit rates and average these over the 2,500 posterior draws (Simoiu, 2018).

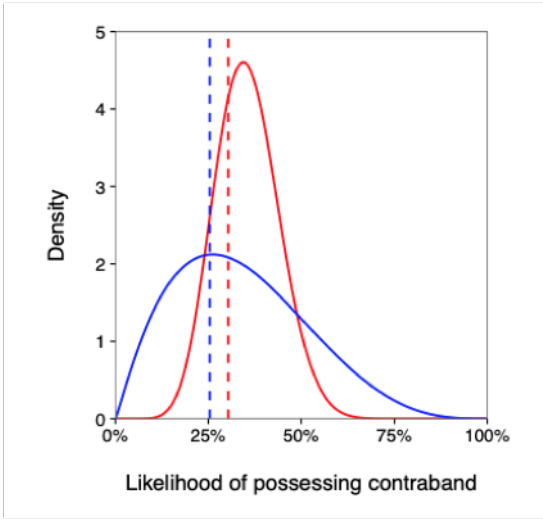


Figure 1: Thresholds of Two Risk Distributions

Raw Data

The publicly available NYPD Stop, Question, and Frisk database contained all the records from stops conducted between 2003 through 2018. The records from 2003 through 2016 were in a .csv format and the records from 2017 through 2018 were in a .xlsx format. For the purposes of consistency we converted these latter datasets into a .csv format. The raw datasets include a multitude of columns valuable to our research, such as if an individual in a stop was searched, if contraband was found, if an arrest was made, and the stop's time and location. The publicly available Historical NYC Crime Data contained the numbers of seven major felony offenses, seven non-major felony offenses, and misdemeanors for each precinct for the years 2000 - 2018. We extracted data from this record from 2003 through 2018.

Methods

Preprocessing of Data and MCMC

The data was first cleaned and filtered using R Studio. The datasets originally presented the details of each police stop, including the race of the suspect, the precinct where the stop was conducted, whether or not the suspect was searched, and whether or not certain types of contraband were found on the individual. From this data, multiple contraband columns were merged to determine whether a search was successful and create a single contraband column. Stops in which the person was not initially searched but had apparent contraband on them were filtered out; in cases like these, the police officer would utilize the visible contraband, not race, as a reason for stopping an individual, and thus, this stop would not indicate . Black Hispanic and white Hispanic individuals were combined to create the single category of Hispanic individuals, while unknown/missing races and Asians, Pacific Islanders, and Native Americans were filtered out. The latter racial groups were not considered because of their smaller demographics, which would cause erroneous statistical analyses. Missing and extraneous data was omitted, and each dataset was converted to find the total stops, total searches, and total hits (successful searches) per race per precinct; in addition, the search rate and hit rate was calculated for each race for each precinct.

The Historical NYC Crime Data was aggregated via Excel to find the total crime numbers (including major felonies, non-major felonies, and misdemeanors) per precinct for the years 2003 to 2018. Through RStudio,

this data was appended to each year's SQF dataset; the SQF datasets were then aggregated as four four-year chunks, including the years 2003 - 2006, 2007 - 2010, 2011 - 2014, and 2015 - 2018.

Next, the threshold test model created by Pierson, Corbett-Davies, and Goel (2018), which uses race and precinct as parameters, was applied to the aggregated data using RStan. The Markov Chain Monte Carlo (MCMC) method was used to infer risk distribution parameters and search thresholds for each race per precinct. The search thresholds for white, black, and Hispanic individuals were then plotted externally against crime numbers to determine if a correlation existed. The discrepancies between minority and white search thresholds were then visualized by comparing the two on a bivariate graph. The parameters and search thresholds were used to find the model-predicted search and hit rates, which were then compared to the actual values in posterior predictive checks to determine the robustness of the model.

```
md_list = list(md1, md2, md3, md4)
stan_data_list = list()

for(i in 1: length(md_list)) {
  stan_data_list[[i]] = with(md_list[[i]], list(
    N = nrow(md_list[[i]]),
    D = length(unique(pct)),
    R = length(unique(race)),
    d = as.integer(pct),
    r = as.integer(race),
    n = numstops,
    s = numsearches,
    h = numhits ))
}

pctNames1 = unique(md1$pct)
pctNames2 = unique(md4$pct)

model <- stan_model(file = 'threshold_old.stan')
```

```
## DIAGNOSTIC(S) FROM PARSER:
## Info (non-fatal): Comments beginning with # are deprecated.  Please use // in place of # for line co
```

```
fit = list()
post = list()
for (i in 1: length(stan_data_list))
{
  fit[[i]] = sampling(
    model, data = stan_data_list[[i]], iter=5000,
    init = 'random', chains=5,
    cores=5, refresh=50, warmup = 2500,
    control = list(adapt_delta = 0.95,
                   max_treedepth = 12,
                   adapt_engaged = TRUE))

  post[[i]] = rstan::extract(fit[[i]])
}
```

Calculating and Visualizing Racial Bias Over Time

From the bivariate graph of search thresholds, we determined the signed distance from each point, representing a precinct, to the dashed line, which represents a state of no racial discrimination. These

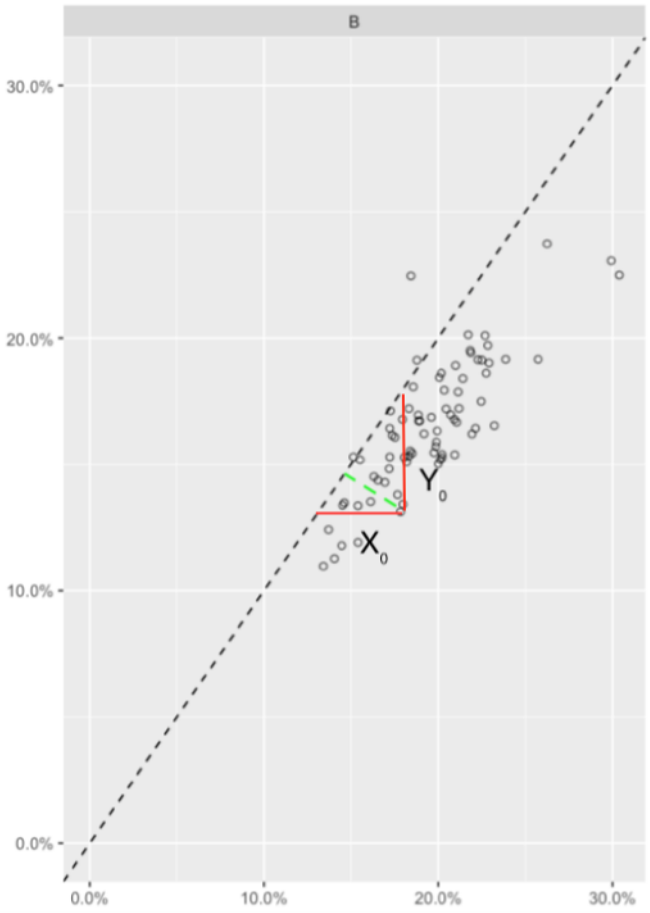


Figure 2: Calculating the racial discrimination index

distances represent the amount of racial bias within the precinct; greater distances suggest more racial bias, while distances closer to zero suggest less racial bias. Positive distances denote discrimination against white individuals, and negative distances signify discrimination against black and Hispanic individuals. These distances thus act as a racial discrimination index (I_D). The racial discrimination index is calculated as follows:

$$I_d = \sqrt{|x_0 - y_0|} \cdot \text{sign}(y_0 - x_0)$$

Using I_D, we then plotted the overall distribution of the amounts of racial discrimination within each precinct as a kernel density function for each four-year chunk to determine overall potential fluctuations of racial bias. The amount of racial bias against black, white, and Hispanic individuals in each precinct was also visualized over time as a heatmap using ArcGIS to portray trends within each precinct. Above is an example of a kernel density distribution plotted over a histogram. The source code for this process can be found below.

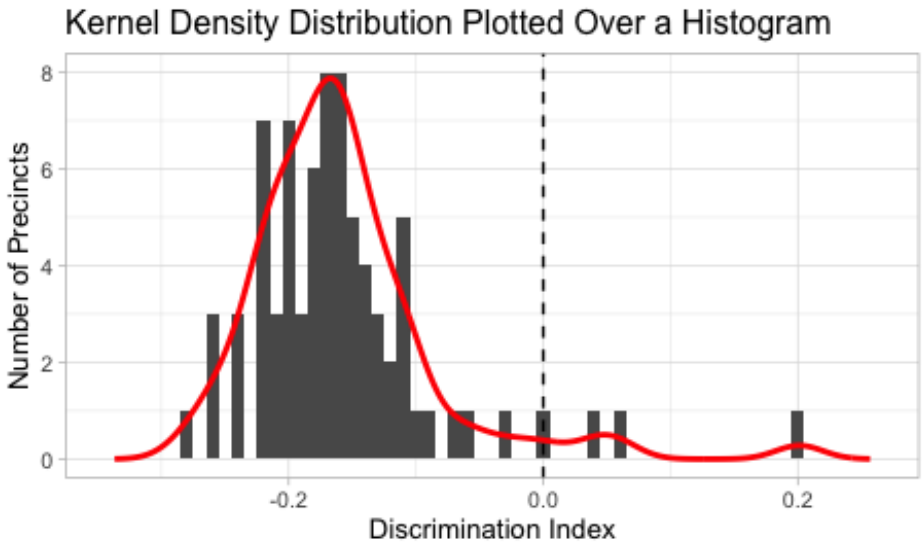


Figure 3: Kernel Desnity Distribution Plotted Over a Histogram

Geocoding and Map Visualizations

The original NYPD SQF dataset was filtered using the same method as previously described; however, the data was left in its original form describing individual stops, rather than aggregating total stops, searches, and hits. The datasets from 2006 - 2018 included a column with x- and y-coordinates from the NAD83 / New York Long Island (ftUS) state plane coordinate system. The script converts these coordinates to latitude and longitude using ArcGIS's geocoding APIs values which can be plotted with a geographic information systems (GIS) software. The datasets from 2003 - 2005 did not include the state plane coordinates. Another python script was written that merged together columns with address information about stops including street numbers, street names, and intersection names. This combined information was fed into the ArcGIS geocoding APIs to generate latitude and longitude coordinates. This process took approximately one week to complete due to the process intensive nature of geocoding addresses. We utilized a high-performance computing cluster to ensure a sturdy network and adequate memory for the task. Through QGIS, each year's stops was visualized as a density heatmap. The heatmaps for all years were then animated as a GIF using ezgif.com. We also utilized ArCGIS to create two map visualizes indicating how racial bias against black and Hispanic individuals has changed over time in the city.

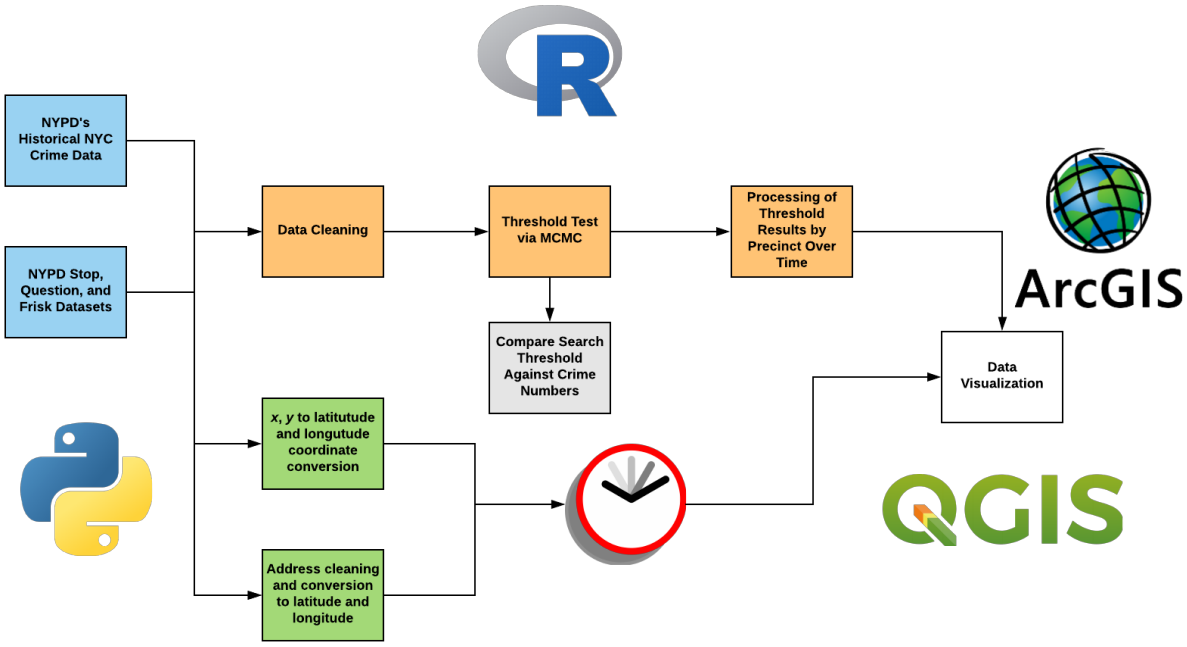
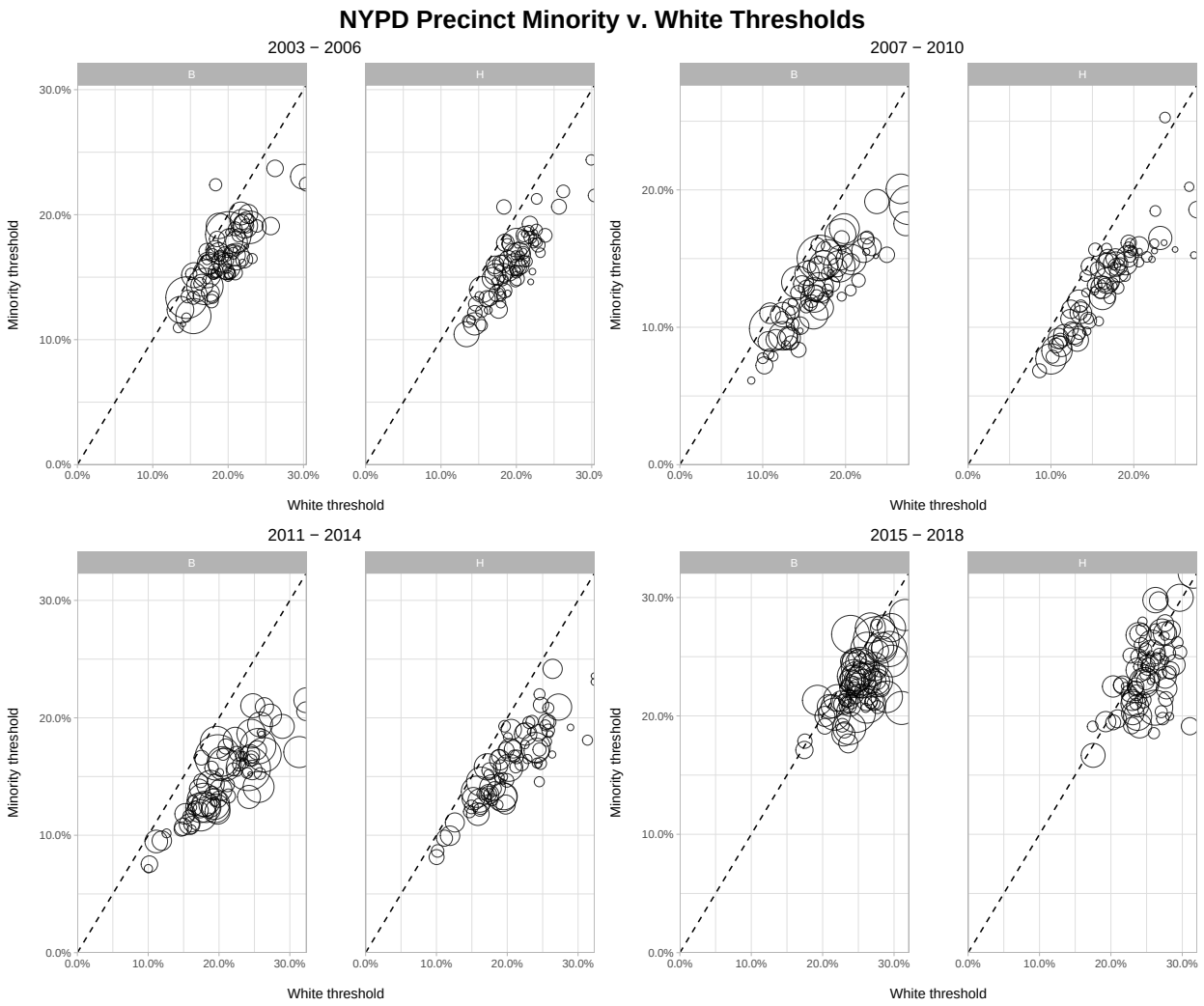


Figure 4: Methodology flowchart for processing SQF datasets

Results

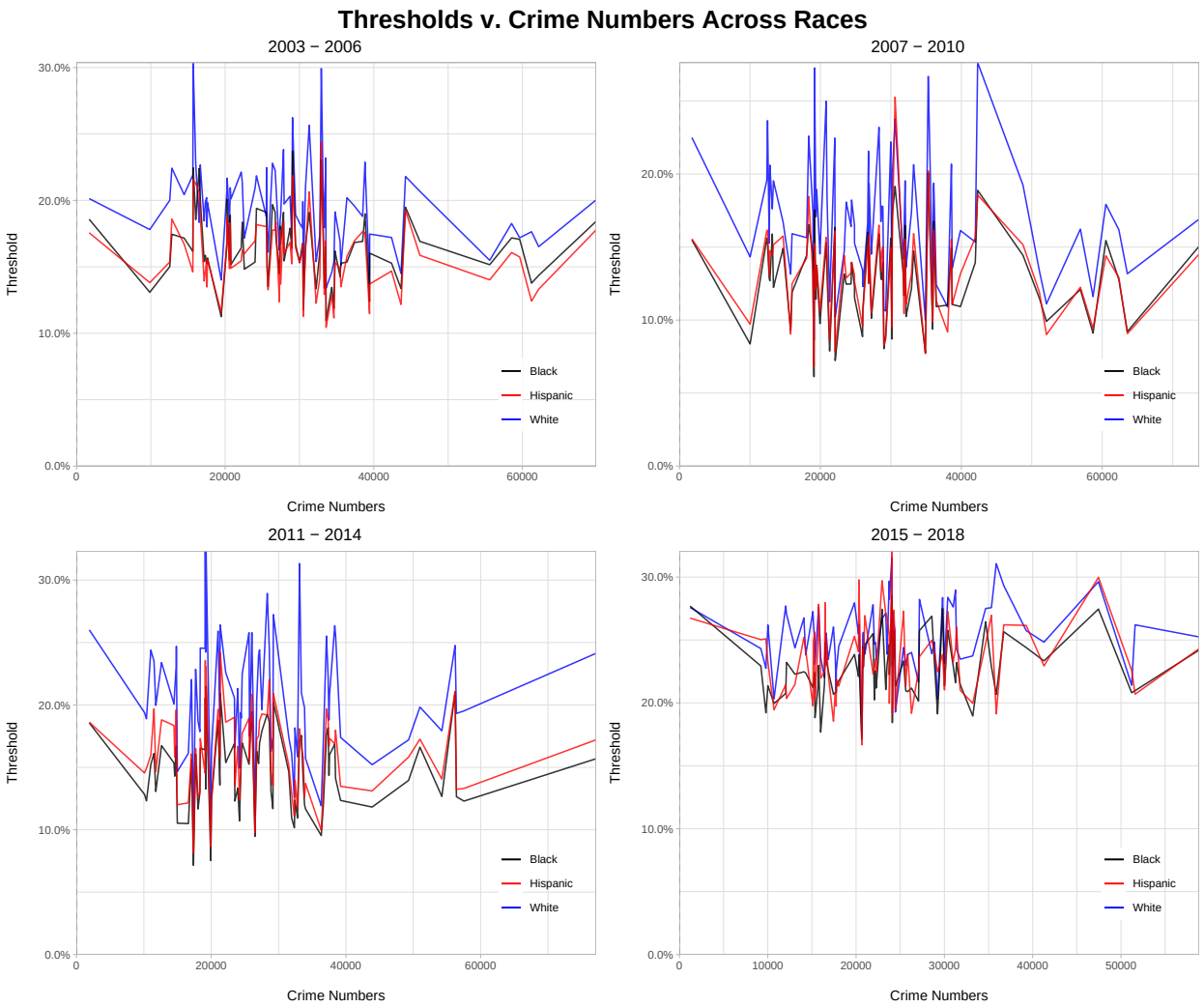
Comparing Minority and White Thresholds

To visualize discriminatory policing practices against black and Hispanic racial groups, minority versus white search thresholds of the various New York City Police precincts are plotted below. The dotted line represents policing practices where no discrimination occuried. Several precincts fall quite close to this line, so it is possible that the regions of these precincts are populated with a particular racial group. The size of the circles represent the number of stops of minorities in each precinct. In the 2015 - 2018 chunk a significant portion of the data is heavily condensed over a higher threshold rate and also closer to the dotted linear regression line.



Search Threshold vs. Crime Number Analysis

The following graphs plots the search threshold changes among white, black, and Hispanic racial groups against the number of total number of crimes in New York City, over four four-year periods from 2003 through 2018. For each of the four-year chunks there is a consistent decrease in thresholds for all racial groups when they are under ten thousand crimes. In the three four-year chunks from 2003 - 2014, the white threshold is almost always consistently higher than the black and Hispanic thresholds regardless of the number of crimes. In the 2015 - 2018 chunk, the thresholds across races are relatively close together as the number of crimes increases. In all chunks, the thresholds for each race consistently vary according to the same pattern, particularly between ten-thousand and forty-thousand crimes. The search thresholds generally increase after sixty-thousand crimes for the three four-year chunks from 2003 - 2014 and after fifty-thousand crimes for the chunk from 2015 -2018.



Average Thresholds and Credibility Intervals

The following table summarizes how the search threshold has changed over four four-year periods from 2003 through 2018. Under all circumstances the threshold for white individuals is higher than that for both black and Hispanic minorities. The average search threshold in the period from 2011 through 2014 is of particular interest. Historically this was at the height of SQF practices by the NYPD and black and Hispanic racial groups are searched at significantly lower search thresholds than their white counterparts. For the three four-year periods from 2003-2011, the 95% credibility interval is typically no larger than a range of 0.3. However, in the last four-year period from 2015 to 2018 this range for the 95% credibility more than doubles. This may be the result of a lack of data points for this last four-year set, an issue discussed in Posterior Predictive Checks.

2003 - 2006

```
##   Driver Race Average Threshold 95% Credible Interval
## 1           B             0.164        (0.151, 0.177)
## 2           H             0.159        (0.145, 0.171)
## 3           W             0.192        (0.175, 0.207)
```

2007 - 2010

```
##   Driver Race Average Threshold 95% Credible Interval
## 1           B             0.130        (0.119, 0.141)
## 2           H             0.132        (0.118, 0.146)
## 3           W             0.168        (0.151, 0.185)
```

2011 - 2014

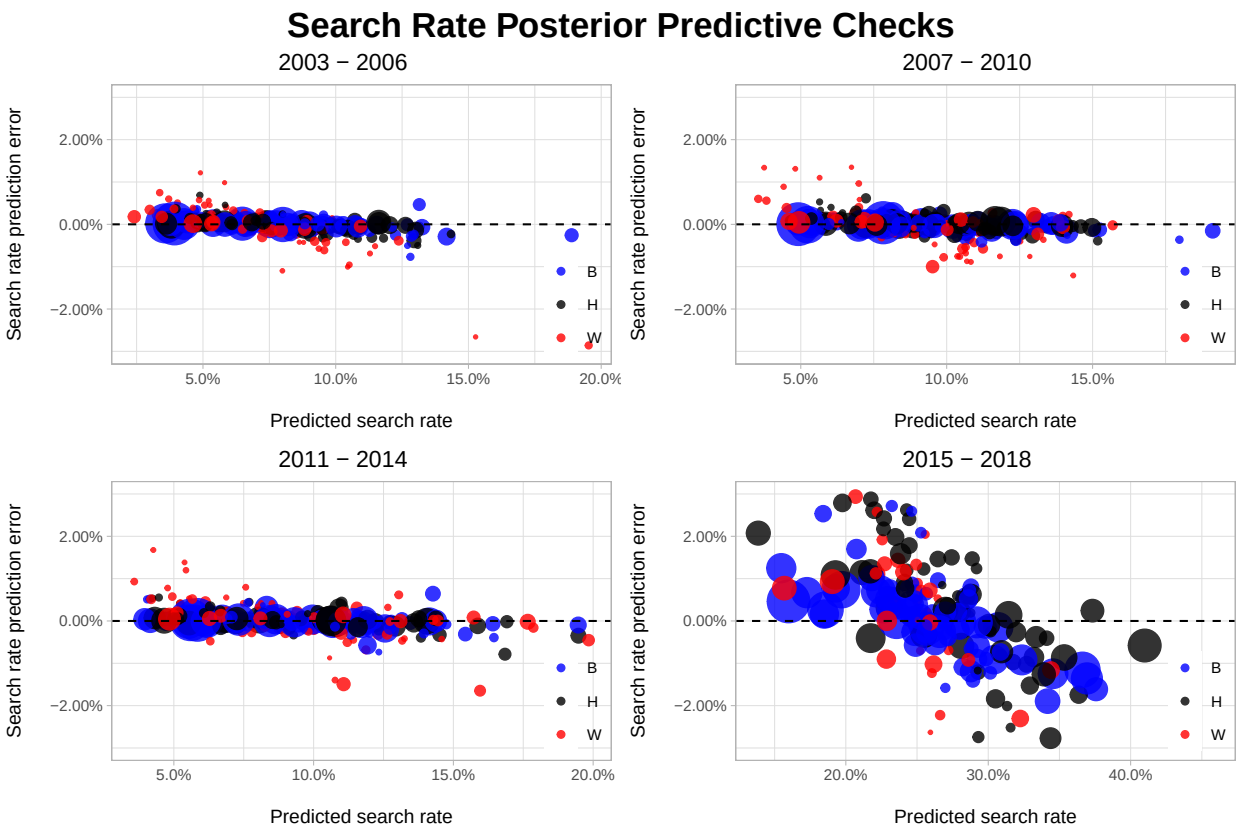
```
##   Driver Race Average Threshold 95% Credible Interval
## 1           B             0.148        (0.136, 0.159)
## 2           H             0.161        (0.148, 0.174)
## 3           W             0.208        (0.191, 0.224)
```

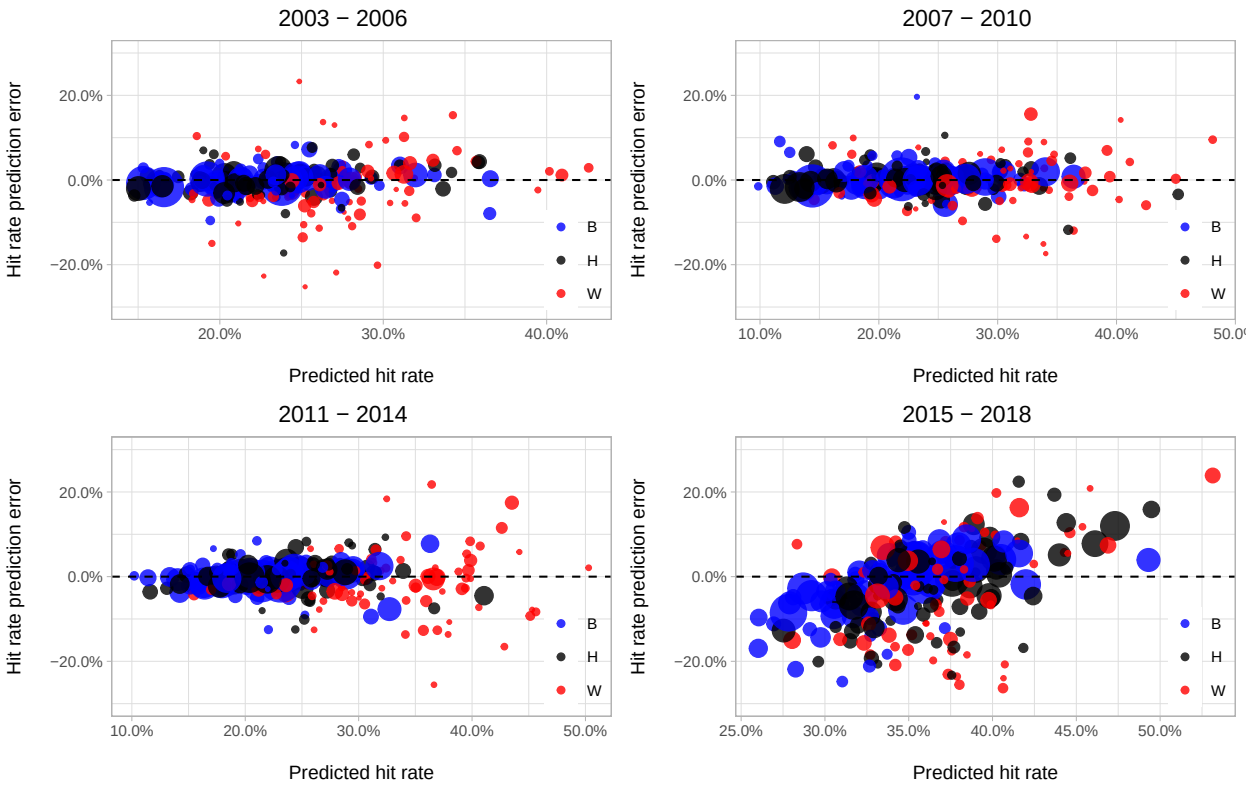
2015 - 2018

```
##   Driver Race Average Threshold 95% Credible Interval
## 1           B             0.228        (0.194, 0.260)
## 2           H             0.238        (0.202, 0.276)
## 3           W             0.254        (0.212, 0.294)
```

Posterior Predictive Checks

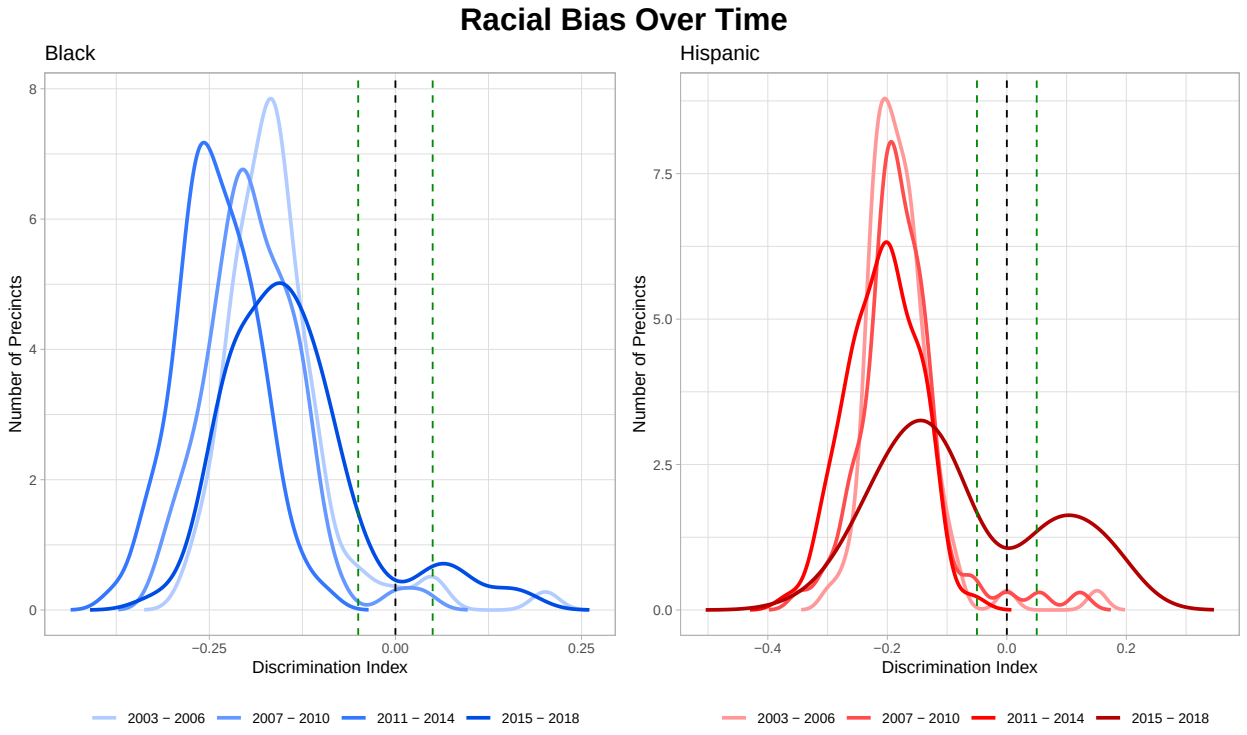
The following graphs present the posterior predictive checks for each of the four-year chunks. For each chunk, the error in the model-predicted search and hit rates per race per precinct were graphed. For the years 2003 - 2014, the search rate error remains relatively small across races and precincts. For the same years, the hit rate error is generally higher with more variance, although it still provides reasonable accuracy for the majority of races and precincts. However, the chunk for years 2015 - 2018 presents large errors in both the model-predicted search and hit rates. This is likely due to the fact that these years had considerably less data than previous years due to a reduction in SQF stops during this time period; after the 2013 court case Floyd v. City of New York debating the constitutionality of the SQF procedure, SQF stops dramatically decreased. These smaller numbers would cause erroneous results when running the threshold test model, which was designed for large-scale statistical analysis. Therefore, the 2015 - 2018 results are inconclusive and may be better studied with a different type of statistical method.


Search Rate Posterior Predictive Checks
2003 − 2006
2007 − 2010
2011 − 2014
2015 − 2018
Search rate prediction error
Predicted search rate
2.00%
0.00%
−2.00%
5.0%
10.0%
15.0%
20.0%
20.0%
30.0%
40.0%
B
H
W


Hit Rate Posterior Predictive Checks
2003 − 2006
2007 − 2010
2011 − 2014
2015 − 2018
Hit rate prediction error
Predicted hit rate
20.0%
0.0%
−20.0%
10.0%
20.0%
30.0%
40.0%
50.0%
25.0%
35.0%
45.0%
B
H
W

Racial Bias Over Time

To visualize fluctuations in racial bias over time, the distribution of racial bias in precincts was plotted for each-four year chunk. Positive discrimination indices indicate bias against white individuals, while negative discrimination indices indicate bias against minorities. In the period from 2003 chunk, black individuals begin with both a higher racial bias index and a greater number of discriminative precincts than those associated with Hispanic individuals, and both minorities are higher in both categories when compared to white individuals. For both black and Hispanic individuals, there is a decrease in both the discrimination index and the number of discriminative precincts over time. This was paired with a simultaneous increase in the number of precincts discriminative toward white individuals, although the discrimination index for white individuals stayed within the same relative range across all the chunks. It is significant to note that racial bias and the number of discriminative precincts for black individuals decreased considerably less than bias against Hispanic individuals. In both graphs, the increase in racial discrimination against whites occurred in a relatively small number of precincts with a relatively small discrimination index in comparison to both black and Hispanic individuals. This implies that racial bias against black individuals has remained more prevalent and more considerable than racial bias against Hispanic individuals measured across time, although both minorities continue to face far greater and far more prevalent discrimination than white individuals across all chunks. The general decreasing trend in racial bias signifies that policing practices are becoming less discriminative.



Map Visualization

The density heatmap of the SQF stops across all years can be seen via this link: https://media.giphy.com/media/XaFRbg4JgbXKxaP00M/giphy.gif. Visualizations like this one grant the ability to determine "hot spots" of NYPD SQF activity throughout the years, which can then be investigated for further analysis.

The ArcGIS heatmap displaying the Hispanic discrimination index for each precinct across the four-year chunks can be accessed with this link: https://njit-r.maps.arcgis.com/apps/TimeAware/index.html?appid=326a003595324aae8fde6f55e4ce8c59

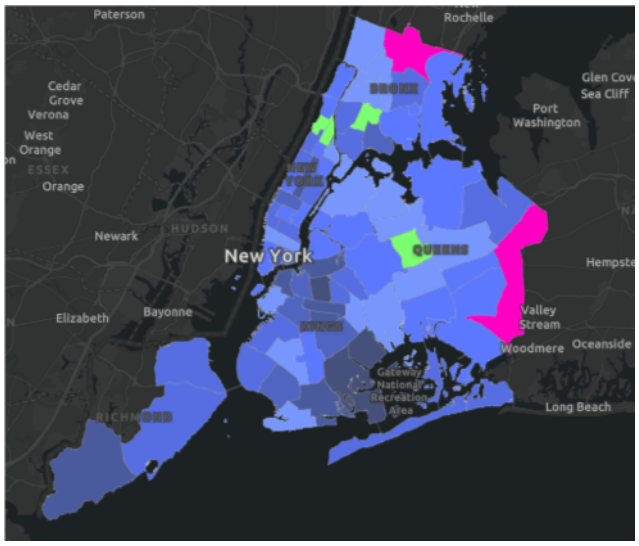


Figure 5: Black Discrimination Index Heatmap

The heatmap displaying the black discrimination index can be accessed with this link: https://njit-r.maps.arcgis.com/apps/TimeAware/index.html?appid=ef6d665772ce49dcb8a55a839c830ff9

Green precincts represent precincts in which the discrimination index for either white or minority groups was less than an absolute value of 0.05; these precincts have little discrimination that may be due to chance rather than implicit racial bias. We consider precincts outside of this range to have discriminatory policing embedded in their SQF practices. Pink precincts represent precincts discriminating against white individuals, while blue and purple precincts discriminating against hispanic and black people, respectively. An example of the Hispanic discrimination index heatmap for a single four-year chunk is shown below.

The ArcGIS heatmaps easily show where racial bias exists within New York City over a 16 year period, allowing policymakers and police chiefs to determine the progression of equality within their precincts and in relation to surrounding precincts to make future policing decisions that work best for their individual communities.

Conclusion

While there was no significant correlation between the search thresholds and neighborhood crime numbers, white search thresholds were consistently higher than black and Hispanic search thresholds, implying consistent racial discrimination against these minorities. However, all thresholds presented near identical patterns of variance across the different crime numbers of separate precincts, implying that different precincts were mostly consistent with how they treated different racial groups.

The model was generally accurate at predicting the search rate of all races but slightly increased in error when predicting the hit rate for the years 2003 - 2014; the large errors in search and hit rates for the years 2015 - 2018 indicate these results are inconclusive due to an insufficiency of data points. This suggests that statistical methods meant for smaller-scale analysis might be better suited to analyzing these years and future years; future studies may consider revising the threshold test model to work with smaller datasets.

Overall, racially discriminative policing against black and Hispanic individuals have slowly decreased across the years 2003 - 2018 in both the prevalence and the extent of discrimination; however, there remains significant work to be done towards eradicating racial bias within the NYPD's SQF procedure.

This information, along with a robust visualization platform for racial bias, provides valuable tools for policymakers, police, and residents to make informed decisions about policing practices in the places that they call home.

Future Work

This research examines the use of the threshold test model within the context of racially-biased policing practices; however, the model has the potential to be applied to a variety of fields, such as banking, college admissions, healthcare, and hiring practices. The model can also be presented in different web application and visualization formats that allow ease of access to both civilians and law enforcement.

Aside from applying the threshold test to different settings, the model itself can be revised to reflect different parameters, such as neighborhood crime or population density, precinct racial demographics, and the time the stop occurred. Additional parameters may increase the accuracy of the model and the credibility of the results. In pursuit of increased accuracy, future studies may also focus on revising the model to create a dual-threshold test, which calculates separate search thresholds for innocent and guilty individuals.

In addition, this research can delve deeper into spatial analysis by adopting spatial weights, which may better present the relationships between biased precincts and neighborhoods. Spatial weights would not only consider stop factors within a single precinct but also consider how the bias in a single precinct would influence its neighbors and vice versa. This type of spatial analysis has the potential to not only help law enforcement but also help urban planners see how bias affects daily life in neighborhoods.

Lastly, we hope this research inspires the creation of objective tools for law enforcement officers and policy-makers to determine the quality of local policing practices in all areas of the United States.

References

Al Baker. “Confronting Implicit Bias in the New York Police Department.” The New York Times, 2018.

Ian Ayres. “Outcome tests of racial disparities in police practices.” Justice Research and Policy, 2002.

A.G. Greenwald, M.R. Banaji. “Implicit social cognition: Attitudes, self-esteem, and stereotypes”. Psychological Review, 1995. Lewis R. Katz. “Terry v. Ohio at Thirty-Five: A Revisionist View.” Mississippi Law Journal, 2004.

Emma Pierson, Sam Corbett-Davies, and Sharad Goel. “Fast threshold tests for detecting discrimination.” Forthcoming, 2018.

Camelia Simoiu, Sam Corbett-Davies, and Sharad Goel. “The problem of infra-marginality in outcome tests for discrimination.” Annals of Applied Statistics, 2017.

Camelia Simoiu, et al. “The threshold test: Testing for racial bias in vehicle searches by police.” StanCon, 2018.

Acknowledgements

- New Jersey Institute of Technology
- The National Science Foundation
- Directors of the Computational Data Analytics for Advancing Human Services REU
  - Dr. Zhi Wei
  - Dr. Lian Duan