A large-scale analysis of racial disparities in police stops across the United States

Emma Pierson*, Camelia Simoiu*, Jan Overgoor*, Sam Corbett-Davies*, Daniel Jenson*, Amy Shoemaker*, Vignesh Ramachandran, Phoebe Barghouty*, Cheryl Phillips*, Ravi Shroff† and Sharad Goel*‡

Stanford Computational Policy Lab
March 13, 2019

EXECUTIVE SUMMARY

To assess racial disparities in police interactions with the public, we compiled and analyzed a dataset detailing nearly 100 million municipal and state patrol traffic stops conducted in dozens of jurisdictions across the country—the largest such effort to date. We analyze these records in three steps. First, we measure potential bias in stop decisions by examining whether black drivers are less likely to be stopped after sunset, when a "veil of darkness" masks one's race. After adjusting for time of day—and leveraging variation in sunset times across the year—we find evidence of bias against black drivers both in highway patrol and in municipal police stops. Second, we investigate potential bias in decisions to search stopped drivers. Examining both the rate at which drivers are searched and the likelihood that searches turn up contraband, we find evidence that the bar for searching black and Hispanic drivers is lower than for searching whites. Finally, we examine the effects of legalizing recreational marijuana on policing in Colorado and Washington state. We find evidence that legalization reduced the total number of searches conducted for both white and minority drivers, but we also find that the bar for searching minority drivers is still lower than for whites post-legalization. We conclude by offering recommendations for improving data collection, analysis, and reporting by law enforcement agencies.

More than 20 million Americans are stopped each year for traffic violations, making this one of the most common ways in which the public interacts with the police [9, 16]. Due to the decentralized nature of policing in the United States—and a corresponding lack of comprehensive and standardized data—it is difficult to rigorously assess the manner and extent to which race plays a role in traffic stops [10]. The most widely cited national statistics come from the Police-Public Contact Survey (PPCS) [9], which is based on a nationally representative sample of approximately 50,000 people who report having been recently stopped by the police. In addition to such survey data, some local and state agencies have released periodic reports on traffic stops in their jurisdictions, and have also made their data available to outside researchers for analysis [2, 3, 7, 14, 21–28]. While useful, these datasets provide only a partial picture. For example, there is concern that the PPCS, like nearly all surveys, suffers from selection bias and recall errors. Data released directly by police departments are potentially more complete, but are available only for select agencies, are typically limited in what is reported, and are inconsistent across jurisdictions.

To address these challenges, we compiled and analyzed a unique dataset detailing nearly 100 million traffic stops carried out by 21 state patrol agencies and 29 municipal police departments over almost a decade. This dataset was built through a series of public records requests filed in all 50 states. To facilitate future analysis, we are redistributing these records in a standardized form. To our knowledge, this is the most comprehensive public release and analysis of U.S. traffic stop records to date.[1]

Our statistical analysis of these records proceeds in three steps. First, we assess potential bias in stop decisions by applying the "veil of darkness" test developed by Grogger and Ridgeway [13]. The test is based on a simple observation: because the sun sets at different times throughout the year, one can examine the racial composition of stopped drivers as a function of sunlight while controlling for time of day. If black drivers make up a smaller share of stopped drivers after sunset, when it is difficult to determine a driver's race, that suggests black drivers were stopped before sunset in part because of their race. In both state patrol and municipal police stops, we find that black drivers comprise a smaller proportion of drivers stopped after sunset, which is suggestive of racial bias in stop decisions.

Second, we investigate potential bias in the post-stop decision to search drivers for contraband. To do so, we apply the *threshold test* recently developed by Simoiu et al. [20, 25]. The threshold test incorporates both the rate

*Stanford University; †New York University
‡Corresponding author: inquiries may be sent to scgoel@stanford.edu.

[1] In an earlier version of this report, our analysis was limited to stops carried out by state patrol agencies, due to data availability. This updated report includes information on municipal police stops, as well as more extensive data on state patrol stops. Moreover, while we previously only considered search decisions, the analysis in this report examines potential bias in both stop decisions and search decisions.

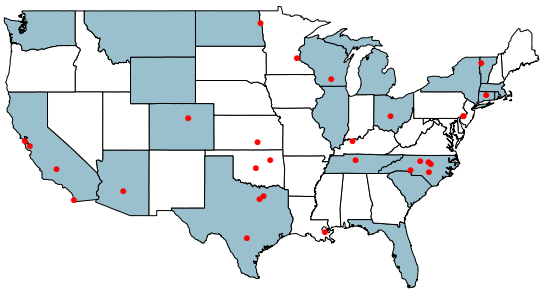

Figure 1: *We collected and analyzed data on approximately 93 million stops from 21 state patrol agencies (blue) and 29 municipal police departments (red) across the country.*

at which searches occur, as well as the success rate of those searches, to infer the standard of evidence applied when determining whom to search. This approach builds on traditional outcome analysis [5, 6], in which a lower search success rate for one group relative to another is seen as evidence of bias against that group, as it suggests a lower evidentiary bar was applied when making search decisions. Applied to our data, the threshold test indicates that black and Hispanic drivers were searched on the basis of less evidence than white drivers, both on the subset of searches carried out by state patrol agencies and on those carried out by municipal police departments.

Finally, we examine the effects of drug policy on traffic stop outcomes. We specifically compare patterns of policing in Colorado and Washington—two states that legalized recreational marijuana at the end of 2012—to 12 states in which recreational marijuana remained illegal. Using a difference-in-differences strategy, we find that legalization reduced both search rates and misdemeanor rates for drug offenses for white, black, and Hispanic drivers—though a gap in search thresholds persists.

In the process of collecting and analyzing millions of traffic stop records across the country, we encountered many logistical and statistical challenges. Based on these experiences, we conclude by offering suggestions to improve data collection, analysis, and reporting by law enforcement agencies. Looking forward, we hope this work provides a road map for measuring racial disparities in policing, and facilitates future empirical research into police practices.

Compiling a national database of traffic stops

Data collection

To assemble a national dataset of traffic stops, we filed public records requests with all 50 state patrol agencies and over 100 municipal police departments. The municipal police departments in this list include those that serve one of the largest 100 cities in the nation, as well as some of the largest cities in each state to achieve geographic coverage.

To date, we have collected (and released) data on approximately 185 million stops carried out by 33 state patrol agencies, and 27 million stops carried out by 49 municipal police departments. In many cases, however, the data we received were insufficient to assess racial disparities (e.g., the race of the stopped driver was not regularly recorded, or only a non-representative subset of stops was provided). For consistency in our analysis, we further restrict to stops occurring in 2011–2017, as many jurisdictions did not provide data on earlier stops. Finally, we limit our analysis to drivers classified as white, black or Hispanic, as there were relatively few recorded stops of drivers in other race groups. Our primary dataset thus consists of approximately 93 million stops from 21 state patrol agencies and 29 municipal police departments, as shown in Figure 1 and described in more detail in Table 1.

Data normalization

As each jurisdiction provided stop data in idiosyncratic formats with varying levels of specificity, we used a variety of automated and manual procedures to create the final dataset. For each recorded stop, we attempted to extract and normalize the date and time of the stop; the county (for state patrol agencies) or police subdivision (e.g., beat, precinct, or zone, for municipal police departments) in which the stop took place; the race, gender, and age of the driver; the stop reason (e.g., speeding); whether a search was conducted; the legal justification for the search (e.g., "probable cause" or "consent"); whether contraband was found during a search; and the stop outcome (e.g., a citation or an arrest). As indicated in Table 1, the information we received varies significantly across states. We therefore restrict each of our specific analyses to the corresponding subset of jurisdictions for which we have the required fields.

In many cases, more than one row in the raw data appeared to refer to the same stop. For example, in several jurisdictions each row in the raw data referred to one *violation*, not one stop. We detected and reconciled such duplicates by matching on a location-specific set of columns. For example, in Colorado we counted two rows as duplicates if they had the same officer identification code, officer first and last name, driver first and last name, driver birth date, stop location (precise to the milepost marker), and stop date and time.

Error correction

The raw data provided to us by state and municipal police agencies often contained errors. We ran numerous automated checks to detect and correct these where possible, although some errors likely remain due to the complex nature of the data. For example, after examining the distribution of recorded values in each jurisdiction, we discovered a spurious density of stops in North Carolina listed as occurring at precisely midnight. As the value "00:00" was likely used to indicate missing information, we treated it as such.

In another example, past work revealed that Texas State Patrol officers incorrectly recorded many Hispanic drivers as white (an error the agency subsequently cor-

	State	City	Stops	Date Range	Date	Time	Geographic Subdivision	Subject Race	Subject Age	Subject Gender	Search Conducted	Contraband Found
1	AZ	Mesa	89,312	2014-2017	•	•		•	•	•		
2	CA	Bakersfield	150,275	2011-2017	•	•	•	•	•	•		
3	CA	San Diego	390,867	2014-2017	•	•	•	•	•	•	•	•
4	CA	San Francisco	476,121	2011-2016	•	•	•	•	•	•	•	•
5	CA	San Jose	94,310	2013-2017	•	•		•			•	•
6	CO	Aurora	170,061	2012-2016	•	•	•	•	•	•		
7	CT	Hartford	18,642	2013-2016	•	•	•	•	•	•	•	•
8	KS	Wichita	444,484	2011-2016	•	•		•	•	•		
9	KY	Owensboro	6,606	2015-2017	•	•	•	•	•	•		
10	LA	New Orleans	252,324	2011-2017	•	•	•	•	•	•	•	•
11	MN	Saint Paul	223,783	2011-2016	•	•	•	•		•	•	
12	NC	Charlotte	625,570	2011-2015	•	•		•	•	•	•	•
13	NC	Durham	136,901	2011-2015	•	•		•	•	•	•	•
14	NC	Fayetteville	222,529	2011-2015	•	•		•	•	•	•	•
15	NC	Greensboro	212,847	2011-2015	•	•		•	•	•	•	•
16	NC	Raleigh	337,390	2011-2015	•	•		•	•	•	•	•
17	ND	Grand Forks	27,480	2011-2016	•	•		•		•		
18	NJ	Camden	139,624	2013-2017	•	•		•	•	•		
19	OH	Cincinnati	119,883	2011-2017	•	•		•		•		
20	OH	Columbus	152,380	2012-2016	•	•	•	•		•	•	
21	OK	Oklahoma City	743,380	2011-2017	•	•	•	•	•	•		
22	OK	Tulsa	451,575	2011-2016	•	•		•		•		
23	PA	Philadelphia	1,090,441	2014-2017	•	•	•	•	•	•	•	•
24	TN	Nashville	2,324,967	2011-2016	•	•	•	•	•	•	•	•
25	TX	Arlington	112,004	2016-2016	•	•	•	•		•	•	
26	TX	Plano	251,733	2012-2015	•	•		•		•	•	
27	TX	San Antonio	1,142,958	2012-2017	•	•	•	•	•	•	•	•
28	VT	Burlington	33,884	2012-2017	•	•		•	•	•	•	•
29	WI	Madison	210,201	2011-2017	•	•	•	•		•		
1	AZ	–	2,200,957	2011-2015	•	•	•	•		•	•	
2	CA	–	24,199,710	2011-2016	•		•	•		•	•	
3	CO	–	2,411,489	2011-2017	•		•	•	•	•	•	•
4	CT	–	445,038	2013-2015	•	•	•	•	•	•	•	•
5	FL	–	3,510,471	2011-2015	•	•	•	•	•	•	•	
6	IL	–	1,569,097	2012-2017	•	•	•	•	•	•	•	•
7	MA	–	1,883,756	2011-2015	•		•	•	•	•	•	
8	MI	–	781,038	2011-2016	•	•	•	•				
9	MT	–	682,388	2011-2016	•	•	•	•	•	•	•	
10	NC	–	3,633,767	2011-2015	•		•	•	•	•	•	•
11	ND	–	266,909	2011-2015	•	•	•	•	•	•		
12	NY	–	6,841,977	2011-2017	•	•	•	•	•	•		
13	OH	–	5,207,761	2011-2015	•	•	•	•		•	•	
14	RI	–	249,183	2011-2015	•	•	•	•	•	•	•	•
15	SC	–	4,337,721	2011-2016	•		•	•	•	•	•	•
16	TN	–	2,012,268	2011-2016	•	•	•	•		•		
17	TX	–	14,812,214	2011-2017	•	•	•	•		•	•	•
18	VT	–	258,806	2011-2015	•	•	•	•	•	•	•	•
19	WA	–	6,257,863	2011-2016	•	•	•	•	•	•	•	•
20	WI	–	1,052,838	2011-2016	•	•	•	•		•	•	•
21	WY	–	172,948	2011-2012	•	•	•	•	•	•		
		Total	**93,440,731**									

Table 1: *Summary of the data for the 29 municipal police departments (top) and 21 state patrol agencies (bottom) used in our analyses. A solid circle signifies that usable data are available for at least two-thirds of stops. Geographic subdivision typically means county (for state patrol agencies) or beat/precinct (for municipal police departments).*

rected).[2] To investigate and adjust for this issue, we imputed Hispanic ethnicity from surnames in the three states for which we have name data: Texas, Arizona, and Colorado.[3] Among drivers with typically Hispanic names, the proportion labeled as Hispanic in the raw data is considerably lower in Texas (37%) than in either Arizona (79%) or Colorado (70%), corroborating past results. Because of this known issue in the Texas data, we re-categorized as "Hispanic" all drivers in Texas with Hispanic names who were originally labeled "white" or who had missing race data; we did not re-categorize drivers in any other jurisdictions.

Our complete data cleaning pipeline is extensive, requiring subjective decisions and thousands of lines of code. For transparency and reproducibility, we have released the raw data, the standardized data, and code to clean and analyze the records at https://openpolicing.stanford.edu.

Assessing bias in traffic stop decisions

Relative to their share of the residential population, we find that black drivers are, on average, stopped more often than whites. In particular, among state patrol stops, the annual per capita stop rate for black drivers is 0.11

[2] http://kxan.com/investigative-story/texas-troopers-ticketing-hispanics-motorists-as-white/

[3] To carry out this imputation, we used a dataset from the U.S. Census Bureau that estimates the racial and ethnic distribution of people with a given surname, for surnames occurring at least 100 times [29]. To increase the matching rate, we performed minor string edits to the names, including removing punctuation and suffixes (e.g., "Jr." and "II"), and considered only the longest word in multi-part surnames. Following previous studies [17, 30], we defined a name as "typically" Hispanic if at least 75% of people with that name identified as Hispanic, and we note that 90% of those with typically Hispanic names identified as Hispanic in the 2000 Census.

compared to 0.08 for white drivers; and among municipal police stops, the annual per capita stop rate for black drivers is 0.23 compared to 0.17 for white drivers.[4] For Hispanic drivers, however, we find stop rates are lower than for whites: 0.05 for stops conducted by state patrol (compared to 0.08 for white drivers), and 0.11 for stops conducted by municipal police departments (compared to 0.17 for white drivers).[5]

These numbers are a starting point for understanding racial disparities in traffic stops, but they do not, in and of themselves, provide evidence of racially disparate treatment. In particular, per capita stop rates do not account for possible race-specific differences in driving behavior, including amount of time spent on the road and adherence to traffic laws. For example, if black drivers, hypothetically, spend more time on the road than whites, that could explain the higher stop rates we see for black drivers, even in the absence of discrimination.

Quantifying potential bias in stop decisions is a statistically challenging problem, in large part because one cannot readily measure the racial distribution of those who actually *violated* traffic laws, as the data only contain information on those *stopped* for such offenses. To mitigate this benchmarking problem, Grogger and Ridgeway [13] proposed a statistical approach known as the "veil of darkness" test. Their method starts from the idea that officers who engage in racial profiling are less able to identify a driver's race after dark than during the day. As a result, if officers are discriminating—all else being equal—one would expect black drivers to make up a smaller share of stopped drivers at night, when a "veil of darkness" masks their race. To account for patterns of driving and police deployment that may vary throughout the day, the test leverages the fact that the sun sets at different times during the year. For example, whereas it is typically dark at 7:00 p.m. during the winter months, it is often light at that time during the summer.

To illustrate the intuition behind this method, in Figure 2 we examine the demographic composition of drivers stopped by the Texas State Patrol at various times of day. Each panel in the plot shows stops occurring in a specific 15-minute window (e.g., 7:00–7:15 p.m.), and the horizontal axis indicates minutes since dusk. Following Grogger and Ridgeway [13], we restrict to white and black drivers—as the ethnicity of Hispanic drivers is not always apparent, even during daylight hours—and we filter out stops that occurred in the approximately 30-minute period between sunset and dusk, when it is neither "light" nor "dark."[6] For each time period, the plot shows a marked drop in the proportion of drivers stopped after dusk who are black, suggestive of discrimination in stop decisions.

We now formally apply the veil-of-darkness test to our entire dataset, fitting the same statistical model described by Grogger and Ridgeway [13]. We specifically fit the following logistic regression model:

$$\Pr(\text{black}_i \mid d_i, t_i, g_i) = \text{logit}^{-1}\left(\beta_d \cdot d_i + \gamma^T \cdot \text{ns}_6(t_i) + \beta_{g[i]}\right),$$

where $\Pr(\text{black} \mid d, t, g)$ is the probability that a stopped driver is black at a certain time t and location g, with darkness status $d \in \{0, 1\}$ indicating whether a stop occurred after dusk ($d = 1$) or before sunset ($d = 0$). In this model, $\text{ns}_6(t)$ is a natural spline over time with six degrees of freedom, and $\beta_{g[i]}$ is a location fixed-effect.[7] The main term of interest is β_d, which describes differences in the composition of stopped drivers between daylight and dark, after adjusting for time and location, with $\beta_d < 0$ suggesting discrimination against black drivers. As recommended by Grogger and Ridgeway [13], we fit the model on stops that occurred during the "inter-twilight period": the range from the earliest time dusk occurs in the year to the latest time dusk occurs in the year. This range is approximately 5 p.m. to 10 p.m., though the precise values differ by location and year. All times in the inter-twilight period are, by definition, light at least once in the year and dark at least once in the year.

We fit the veil-of-darkness model separately on the subset of stops carried out by state patrol agencies and on those carried out by municipal police departments. We also fit variations of the basic model in which we: (1) altered the granularity of the location fixed-effects (e.g., city vs. precinct); and (2) added interaction terms between time and location. We also considered splines with different degrees of freedom, but those variants yielded nearly identical results and so are not discussed further.

Data	Controls	β_d	s.e.
All states	time + state	-0.116	0.0052
	time × state	-0.106	0.0052
	time + county	-0.105	0.0054
	time × county	-0.104	0.0055
All cities	time + city	-0.024	0.0051
	time × city	-0.024	0.0052
Cities with sub-geography	time + city	-0.048	0.0062
	time × city	-0.045	0.0063
	time + sub-geography	-0.058	0.0068
	time × sub-geography	-0.056	0.0069

Table 2: *Results from the veil-of-darkness test to assess bias in stop decisions. After adjusting for time of day, black drivers comprise a smaller share of stopped drivers after dark (indicated by negative values of β_d), suggestive of racial profiling. These results are consistent across different subsets of the data and under different model specifications.*

[4] These numbers are the unweighted average annual per capita stop rates across the jurisdictions we analyze.

[5] Hispanic drivers also report being stopped less often that white drivers in the Police-Public Contact Survey [9].

[6] We distinguish here between *sunset* and *dusk*. Sunset is the point in time where the sun dips below the horizon. Dusk, also known as the end of civil twilight, is the time when the sun is six degrees below the horizon, and when it is widely considered to be "dark."

[7] For computational efficiency, time is rounded to the nearest 5-minute interval when fitting the model. Note that $g[i]$ is the location for stop i, and $\beta_{g[i]}$ the corresponding coefficient.

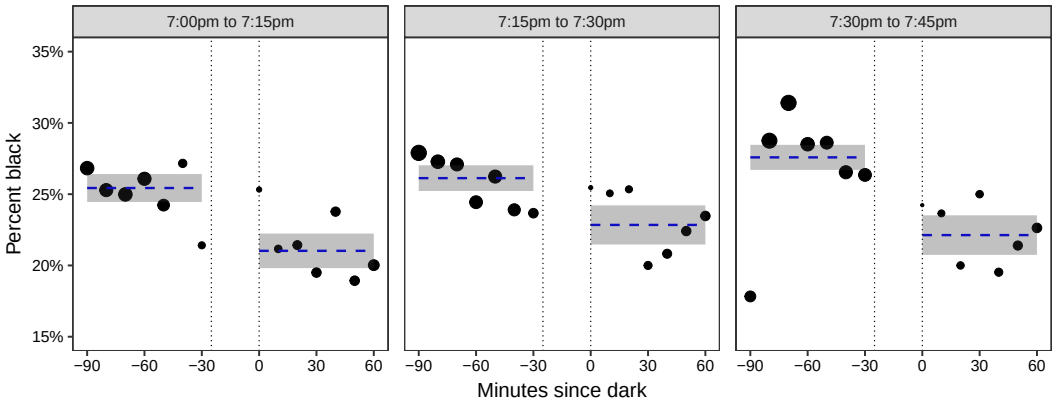

Figure 2: *As an illustration of the veil-of-darkness test, we consider stops carried out in narrow time windows in a single state (Texas), and compute the percentage of stops that involved black drivers for a series of 10-minute bins before and after dusk, among stops of black and white drivers. The vertical line at $t = 0$ indicates dusk, at which point it is generally considered "dark"; we remove stops in the approximately 30-minute period between sunset (indicated by the left-most vertical line) and dusk, as this period is neither "light" nor "dark." The dashed horizontal lines show the overall share of stops involving black drivers before and after dark, and the gray bands indicate 95% confidence intervals. For all three depicted time windows, black drivers comprise a smaller share of stopped drivers after dark, when a "veil of darkness" masks their race, suggestive of racial profiling.*

Our results are summarized in Table 2. Across the ten different model specifications, we find β_d is consistently negative—and statistically significant—both for stops by state patrol agencies and for those by municipal police departments. In magnitude, β_d ranges from -0.024 (for a model fit on city-level data without sub-geographic controls) to -0.12 (for a model fit on the state-level data with state fixed-effects). These findings are broadly suggestive of racial discrimination against black drivers in stop decisions.

The veil-of-darkness test is a popular technique for assessing disparate treatment, but, like all statistical methods, it comes with caveats. Perhaps most importantly, darkness—after adjusting for time of day—is a function of the date. As such, to the extent that driver behavior changes throughout the year, and these changes are correlated with race, the test can suggest discrimination where there is none. Likewise, if driving behavior is more related to lighting than time of day, that could similarly skew the results. Conversely, artificial lighting (e.g., from street lamps) can weaken the relationship between sunlight and visibility, and so the method may underestimate the extent to which stops are predicated on perceived race. Nevertheless, despite these shortcomings, we believe the test provides a useful if imperfect measure of bias in stop decisions.

Assessing bias in search decisions

After stopping a driver, officers may carry out a search of the driver or vehicle if they suspect more serious criminal activity. We next investigate potential bias in these search decisions.

Among stopped drivers, we find that blacks and Hispanics were, on average, searched more often than whites. Specifically, in the 16 state patrol agencies for which we have the necessary data, search rates were 3.8%, 3.6%, and 1.6% for stopped black, Hispanic, and white drivers, respectively.[8] The analogous numbers for the 18 municipal police departments in which we have data are 15%, 13%, and 11% for black, Hispanic and white drivers, respectively. However, as with differences in stop rates, the disparities we see in search rates are not necessarily the product of discrimination. Minority drivers might, hypothetically, carry contraband at higher rates than whites, and so elevated search rates may result from routine police work even if no racial bias were present.

To measure the role of race in search decisions, we apply two statistical strategies: outcomes analysis and threshold analysis. To do so, we limit to the nine state patrol agencies and six municipal police departments for which we have detailed data on the location of stops, whether a search occurred, and whether those searches yielded contraband.[9]

The outcome test

We start with the *outcome test*, originally proposed by Becker [5, 6] to circumvent omitted variable bias in traditional tests of discrimination. The outcome test is based not on the search rate, but on the *hit rate*: the proportion of searches that successfully turn up contraband. Becker argued that even if minority drivers are more likely to carry contraband, absent discrimination, searched mi-

[8] As before, these numbers are the unweighted average search rates across jurisdictions.

[9] We specifically consider state patrol agencies in Colorado, Connecticut, Illinois, North Carolina, Rhode Island, South Carolina, Texas, Washington, and Wisconsin; and municipal police departments in San Diego, San Francisco, New Orleans, Philadelphia, Nashville, and San Antonio. We defer to each department's characterization of "contraband" when carrying out this analysis.

norities should still be found to have contraband at the same rate as searched whites. If searches of minorities are less often successful than searches of whites, it suggests that officers are applying a double standard, searching minorities on the basis of less evidence.

In Figure 3 (top row), we plot hit rates by race and location for the states (left, Figure 3a) and for the cities (right, Figure 3b) for which we have the necessary information. Across jurisdictions, we consistently see that searches of Hispanic drivers are less successful than those of white drivers. However, searches of white and black drivers have more comparable hit rates. Aggregating across state patrol stops, searches of Hispanic drivers yielded contraband 26% of the time, compared to 36% for searches of white drivers and 32% for searches of black drivers. Similarly, aggregating across municipal police departments, searches of Hispanic drivers yielded contraband 13% of the time, compared to 20% for searches of white drivers and 17% for searches of black drivers.[10] The outcome test thus indicates that search decisions may be biased against Hispanic drivers, but the evidence is more ambiguous for black drivers.

The threshold test

The outcome test is intuitively appealing, but it is not a perfect barometer of bias; in particular, it suffers from the problem of *infra-marginality* [3, 4]. To illustrate this shortcoming, suppose that there are two, easily distinguishable types of white drivers: those who have a 5% chance of carrying contraband, and those who have a 75% chance of carrying contraband. Likewise assume that black drivers have either a 5% or 50% chance of carrying contraband. If officers search drivers who are at least 10% likely to be carrying contraband, then searches of whites will be successful 75% of the time whereas searches of blacks will be successful only 50% of the time. Thus, although the search criterion is applied in a race-neutral manner, the hit rate for blacks is lower than the hit rate for whites, and the outcome test would (incorrectly) conclude searches are biased against black drivers. The outcome test can similarly fail to detect discrimination when it is present.

To mitigate this limitation of outcome tests, the *threshold test* has been proposed as a more robust means for detecting discrimination [20, 25]. This test aims to estimate race-specific probability thresholds above which officers search drivers—for example, the 10% threshold in the hypothetical situation above. Even if two race groups have the same observed hit rate, the threshold test may find that one group is searched on the basis of less evidence, indicative of discrimination.

The threshold test is based on a stylized Bayesian model of officer behavior. During each stop, officers observe a myriad of contextual factors—including the age and gender of the driver, the stop time and location, and behavioral indicators of nervousness or evasiveness. The test assumes that officers distill these factors down to a single number that represents the likelihood the driver is carrying contraband, and that officers conduct a search if that probability exceeds a fixed race- and location-specific threshold. Since there is variation in who is pulled over in any given stop, the probability of finding contraband is modeled as a random draw from a race- and location-specific *signal* distribution. The threshold test then jointly estimates these search thresholds and signal distributions. In this framing, lower search thresholds for one group relative to another are interpreted as evidence of bias.[11]

As shown in Figure 3 (bottom row), the threshold test indicates that the bar for searching black and Hispanic drivers is generally lower than for searching white drivers across the locations we consider. In aggregate across cities, the inferred threshold for white drivers is 15%, compared to 11% for blacks and 10% for Hispanics.[12] These estimated gaps in search thresholds between whites and minorities are large and statistically significant: the 95% credible interval for the white-Hispanic difference is (4%, 6%); and the corresponding interval for the white-black difference is (3%, 6%). Similarly across states, the inferred threshold for white drivers is 24%, compared to 18% for blacks and 15% for Hispanics. These differences are again large and statistically significant: the 95% credible interval for the white-Hispanic gap is (7%, 10%); and the analogous interval for the white-black gap is (4%, 7%).

Whereas the by-location hit rates from the outcome test indicate discrimination only against Hispanic drivers, the threshold test suggests discrimination against both blacks and Hispanics. Consistent with past work [25], this difference appears to be driven by a small but disproportionate number of black drivers who have high inferred likelihood of carrying contraband. Thus, even though the threshold test finds the bar for searching black drivers is lower than for whites, these groups have similar hit rates.

The threshold test provides evidence of racial bias in search decisions. However, as with all tests of discrimination, it is important to acknowledge limits in what one can conclude from such statistical analysis alone. For example, if search policies differ not only across but also within the geographic subdivisions we consider, then the threshold test might mistakenly indicate discrimination where there is none. Additionally, if officers disproportionately suspect more serious criminal activity when searching black and Hispanic drivers compared to whites (e.g., possession of larger quantities of contraband), then lower observed thresholds may stem from non-discriminatory police practices.

The effects of legalizing recreational marijuana on traffic stop outcomes

We conclude our analysis by investigating the effects of legalizing recreational marijuana on stop outcomes. We specifically examine Colorado and Washington, two states in which marijuana was recently legalized and for

[10] These numbers indicate unweighted averages across states and cities, respectively.

[11] In our analysis, we apply a computationally fast variant of the threshold test proposed by Pierson et al. [20], which we fit separately on state patrol and municipal police stops.

[12] As with our outcome results, these aggregate thresholds are computed by taking an unweighted average of city-specific thresholds; below we similarly report unweighted averages across state-specific thresholds.

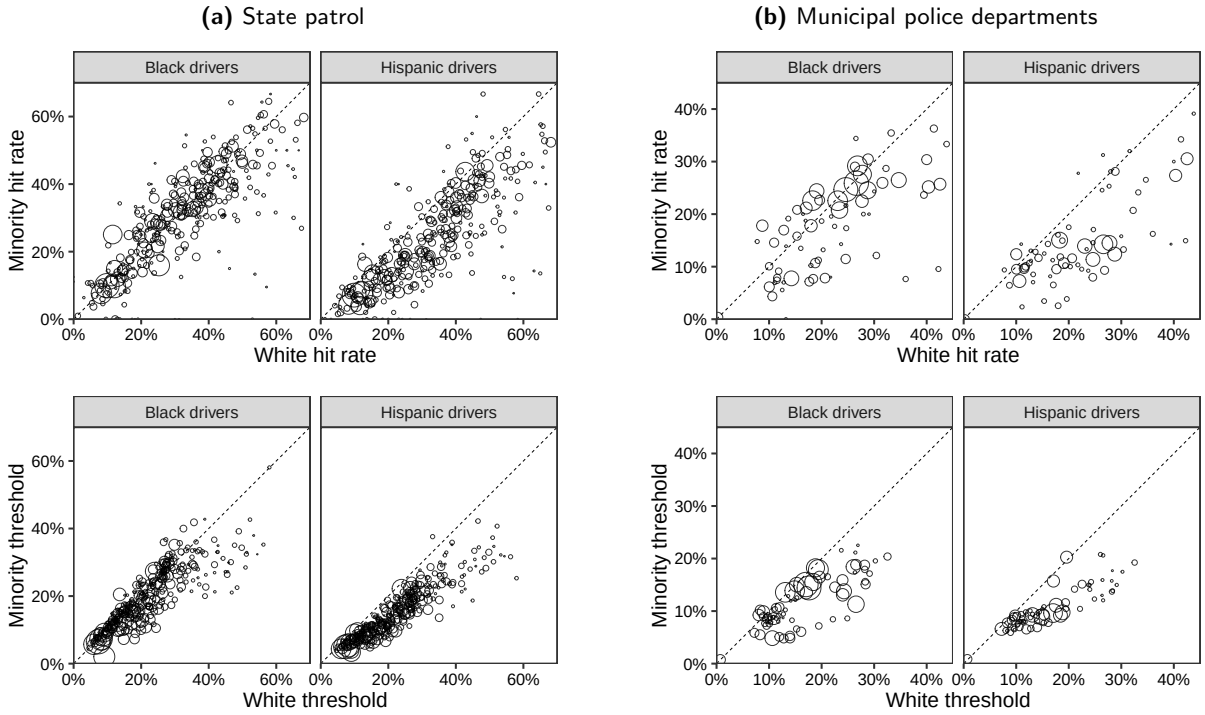

Figure 3: *Hit rates (top) and inferred search thresholds (bottom) by race and location. Plots in the left column (3a) represent 446,000 state patrol searches in nine states, with points corresponding to counties. Plots in the right column (3b) represent 215,000 municipal police searches in six cities, with points corresponding to police precincts. Points are sized by number of searches. Across locations, the inferred thresholds for searching black and Hispanic drivers are typically lower than those for searching white drivers, suggestive of bias. Despite these lower inferred search thresholds, hit rates for blacks are comparable to hit rates for whites, possibly due to the problem of infra-marginality in outcome tests.*

which we have detailed data before and after legalization. As shown in Figure 4 (top), the proportion of stops that resulted in either a drug-related infraction or a drug-related misdemeanor fell substantially in both states after marijuana was legalized at the end of 2012, in line with expectations.[13] In Colorado, we consider only offenses for marijuana possession; in Washington, we include all drug-related misdemeanors, as more detailed information is not available, and so there are still some recorded drug violations post-legalization. Notably, since black drivers were more likely to be charged with such offenses prior to legalization, black drivers were also disproportionately impacted by the policy change. This finding is consistent with past work showing that marijuana laws disproportionately affect minorities [18].

Because the policy change decriminalized an entire class of behavior (i.e., possession of minor amounts of marijuana), it is not surprising that drug offenses correspondingly decreased. However, it is less clear, a priori, how the change might affect officer behavior more broadly. Investigating this issue, we find that after marijuana was legalized, the number of searches fell substantially in Colorado and Washington (Figure 4, bottom), ostensibly because the policy change removed a common reason for conducting searches.[14]

Since black and Hispanic drivers were more likely to be searched prior to legalization, the policy change reduced the absolute gap in search rates between white and minority drivers; however, the relative gap persists, with minorities still more likely to be searched than whites. We further note that marijuana legalization has secondary impacts for law-abiding drivers, as fewer searches overall means fewer searches of those without contraband. In the year after legalization in Colorado and Washington, 40% fewer drivers were searched with no contraband found than in the year before legalization.

As shown in Figure 5, in the twelve states where marijuana was not legalized—and for which we have the necessary search data—search rates did not drop significantly at the end of 2012. This pattern further suggests that the observed drop in search rates in Colorado and Washington is due to marijuana legalization. To add quantitative detail to this visual result, we compute a simple difference-in-differences estimate [1]. Specifically, we fit the following search model on the set of stops in the 14

[13] In these plots, we exclude data for the fourth quarter of 2012, since that period includes stops both before and after legalization.

[14] In both states, we exclude searches incident to an arrest and other searches that are conducted as a procedural matter, irrespective of any suspicion of drug possession.

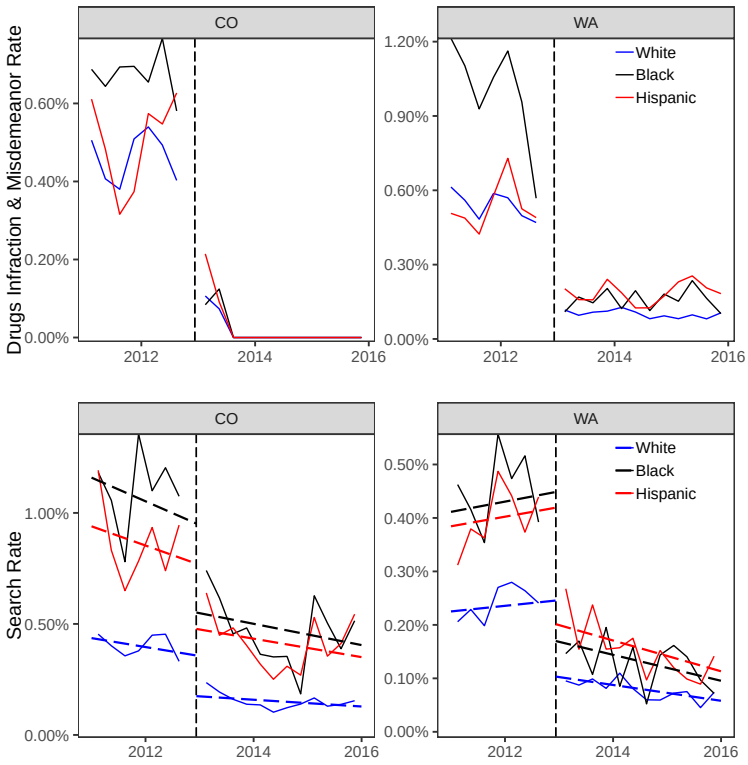

Figure 4: *The proportion of stops that result in a drug-related infraction or misdemeanor (top) or search (bottom) before and after recreational marijuana was legalized in Colorado and Washington at the end of 2012 (indicated by the vertical lines). Subsequent to legalization, there is a substantial drop in offense and search rates. The dashed lines show fitted linear trends pre- and post-legalization.*

states we consider here (Colorado, Washington, and the twelve non-legalization states in Figure 5):

$$\Pr(Y_i = 1) = \text{logit}^{-1}\left(\beta^{\text{state}}_{s[i]} + \beta^{\text{race}}_{r[i]} + \beta^{\text{time}} t_i + \alpha^{\text{race}}_{r[i]} Z_i\right),$$

where Y indicates whether a search was conducted, β^{state}_s and β^{race}_r are state and race fixed-effects, and β^{time} is a time trend, with t a continuous variable in units of years since legalization (e.g., $t = 0.5$ means 6 months post-legalization). The Z term indicates "treatment" status; that is, $Z_i = 1$ in Colorado and Washington for stops carried out during the post-legalization period, and $Z_i = 0$ otherwise. Thus the key parameters of interest are the race-specific treatment effects α^{race}_r. Table 3 lists coefficients for the fitted model. We find that α^{race}_r is large and negative for whites, blacks, and Hispanics, which again suggests the observed drop in searches in Colorado and Washington was due to the legalization of marijuana in those states.

Despite marijuana legalization decreasing search rates for these three race groups, Figure 4 shows that the relative disparity between whites and minorities remains. We apply the threshold test to assess the extent to which this disparity in search rates may reflect bias. Examining the inferred thresholds (shown in Figure 6), we see that white drivers face consistently higher search thresholds than minority drivers, both before and after marijuana legalization. The data thus suggest that although overall search rates dropped in Washington and Colorado, black and Hispanic drivers still face discrimination in search decisions.

Figure 6 also shows that the median threshold faced by all groups decreases after legalization (though not all drops are statistically significant). There are several pos-

	Coef.	s.e.
Effect of legalization on white drivers	-1.00	0.02
Effect of legalization on black drivers	-0.98	0.06
Effect of legalization on Hispanic drivers	-0.79	0.03
Time (years)	-0.02	0.00
Black driver	0.75	0.00
Hispanic driver	0.65	0.00

Table 3: *Effects of legalizing recreational marijuana on search rates, as estimated with a difference-in-difference model. All race groups experienced a large drop in search rate.*

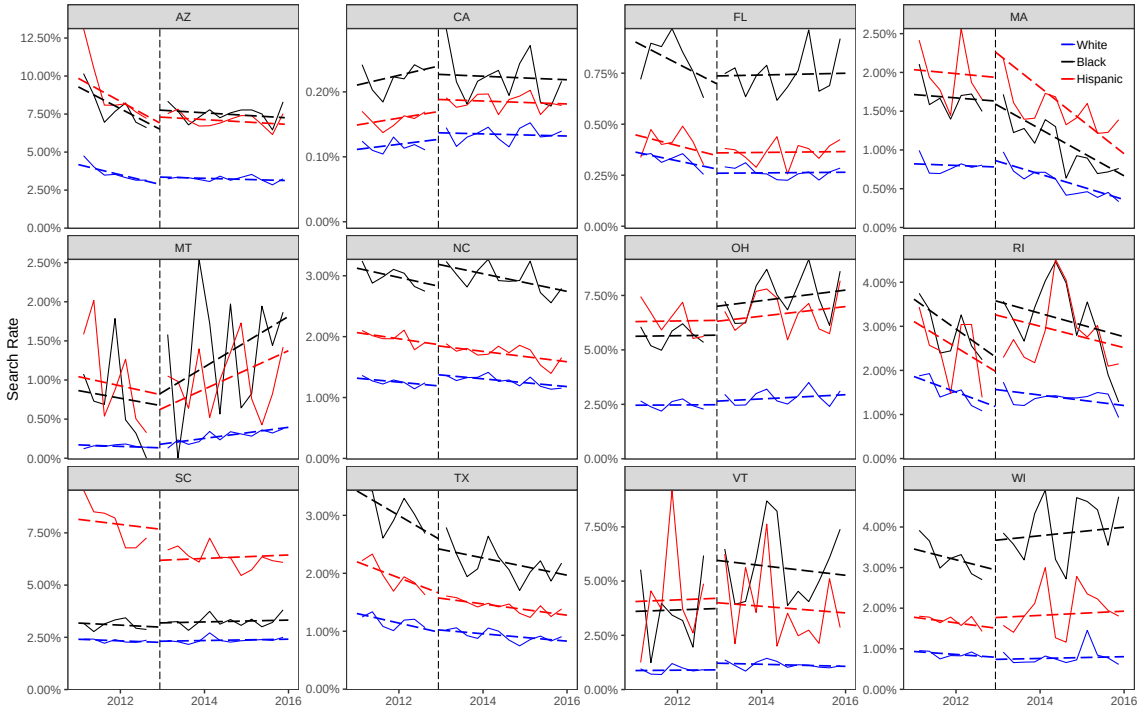

Figure 5: *In the twelve states where marijuana was not legalized, and for which we have the necessary search data, search rates do not fall at the end of 2012; this pattern further suggests that marijuana legalization caused the observed drop in search rates in Colorado and Washington.*

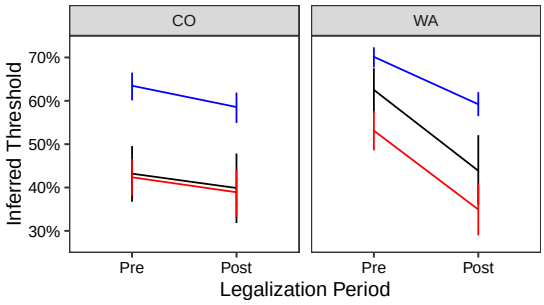

Figure 6: *Inferred median thresholds faced by white (blue lines), black (black lines), and Hispanic (red lines) drivers before and after marijuana legalization. Error bars show the 95% credible intervals of the posterior thresholds. In all cases minority drivers face a lower threshold than white drivers.*

sible explanations for this decrease. Officers may not have fully internalized the change of policy, searching people who would have been at risk of carrying contraband before legalization, but are no longer high risk now that marijuana is legal. Alternatively, or in addition, officers may now be focused on more serious offenses (such as drug trafficking), applying a lower threshold commensurate with the increase in the severity of the suspected crime. Finally, officers may have more resources after being relieved of the task of policing marijuana possession, freeing them to make searches with a lower chance of finding contraband.

Discussion

Our investigation of nearly 100 million traffic stops across the United States reveals evidence of widespread discrimination in decisions to stop and search drivers. Moreover, our analysis of one specific policy change—legalization of recreational marijuana—indicates that such laws can have significant and unexpected downstream consequences on police behavior. In aggregate, our results lend insight into the differential impact of policing on minority communities on an unprecedented scale.

Our study provides a unique perspective on working with large-scale policing data. We conclude by offering several recommendations for data collection, release, and analysis. At minimum, we encourage jurisdictions to collect individual-level stop data that include the date and time of the stop; the location of the stop; the race, gender, and age of the driver; the stop reason; whether a search was conducted; the search type (e.g., "probable cause" or "consent"); whether contraband was found during a search; the stop outcome (e.g., a citation or an arrest); and the specific violation the driver was charged with. Most jurisdictions collect only a subset of this information. There are also variables that are currently rarely

collected but would be useful for analysis, such as indicia of criminal behavior, an officer's rationale for conducting a search, and short narratives written by officers describing the incident. New York City's UF-250 form for pedestrian stops is an example of how such information can be efficiently collected [12, 19].

Equally important to data collection is ensuring the integrity of the recorded information. We frequently encountered missing values and errors in the data (e.g., implausible values for a driver's age and invalid racial categorizations). Automated procedures can be put in place to help detect and correct such problems. In most cases, the recorded race of the driver is based on the officer's perception, rather than a driver's self-categorization. While there are sound reasons for this practice, it increases the likelihood of errors, a problem we observed in the Texas State Patrol data. To quantify and correct for this issue, police departments might regularly audit their data, possibly by comparing an officer's perception of race to a third party's judgment based on driver's license photos for a random sample of stopped drivers.

Despite the existence of public records laws, several jurisdictions failed to respond to our repeated requests for information. We hope law enforcement agencies consider taking steps to make data more accessible to external researchers and to the public. Connecticut and North Carolina are at the forefront of opening up their data, providing online portals for anyone to download and analyze this information.

Finally, we hope that police departments start regularly analyzing their data and report the results of their findings. Such analyses might include estimates of stop, search, and hit rates, stratified by race, age, gender, and location; distribution of stop reasons by race; and trends over time. More ambitiously, departments could use their data to design statistically informed guidelines that encourage more consistent, efficient, and equitable decisions [8, 11, 12, 15]. Many of these analyses can be automated and re-run regularly with little marginal effort. In conjunction with releasing the data underlying these analyses, we recommend the analysis code also be released to ensure reproducibility. Collecting, releasing, and analyzing police data are essential steps for increasing the effectiveness and equity of law enforcement practices, and for improving relations with the public through transparency.

Acknowledgements

We thank Bethney Bonilla, Walter Kim, Joe Nudell, Samantha Robertson, and Eric Sagara for their assistance throughout this project; we also thank Alex Chohlas-Wood and Avi Feller for their helpful feedback. This work was supported in part by the John S. and James L. Knight Foundation and by the Hellman Foundation. All data and analysis scripts are available online at: https://openpolicing.stanford.edu.

References

1. Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008.
2. Kate Antonovics and Brian Knight. A new look at racial profiling: Evidence from the Boston police department. *The Review of Economics and Statistics*, 91(1):163–177, 2009.
3. Shamena Anwar and Hanming Fang. An alternative test of racial prejudice in motor vehicle searches: Theory and evidence. *The American Economic Review*, 2006.
4. Ian Ayres. Outcome tests of racial disparities in police practices. *Justice Research and Policy*, 4(1-2):131–142, 2002.
5. Gary Becker. *The economics of discrimination*. University of Chicago Press, 1957.
6. Gary Becker. Nobel lecture: The economic way of looking at behavior. *Journal of Political Economy*, pages 385–409, 1993.
7. Alex Chohlas-Wood, Sharad Goel, Amy Shoemaker, and Ravi Shroff. An analysis of the Metropolitan Nashville Police Department's traffic stop practices. Technical report, Stanford Computational Policy Lab, 2018.
8. Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017.
9. Elizabeth Davis, Anthony Whyde, and Lynn Langton. Contacts between police and the public, 2015. Technical report, Bureau of Justice Statistics, 2018.
10. Charles Epp, Steven Maynard-Moody, and Donald Haider-Markel. *Pulled over: How police stops define race and citizenship*. University of Chicago Press, 2014.
11. Sharad Goel, Justin Rao, and Ravi Shroff. Personalized risk assessments in the criminal justice system. *The American Economic Review*, 106(5):119–123, 2016.
12. Sharad Goel, Justin Rao, and Ravi Shroff. Precinct or prejudice? Understanding racial disparities in New York City's stop-and-frisk policy. *Annals of Applied Statistics*, 2016.
13. Jeffrey Grogger and Greg Ridgeway. Testing for racial profiling in traffic stops from behind a veil of darkness. *Journal of the American Statistical Association*, 101(475):878–887, 2006.
14. Rebecca Hetey, Benoît Monin, Amrita Maitreyi, and Jennifer Eberhardt. Data for change: A statistical analysis of police stops, searches, handcuffings, and arrests in oakland, calif., 2013-2014. Technical report, Stanford University, SPARQ: Social Psychological Answers to Real-World Questions, 2016.
15. Jongbin Jung, Connor Concannon, Ravi Shroff, Sharad Goel, and Daniel G. Goldstein. Simple rules for complex decisions. Working paper, 2017.
16. Lynn Langton and Matthew Durose. Police behavior during traffic and street stops, 2011. Technical report, U.S. Department of Justice, 2013.
17. Melendres v. Arpaio. Melendres v. Arpaio, 2009. 598 F. Supp. 2d 1025 (D. Ariz. 2009).
18. Ojmarrh Mitchell and Michael Caudy. Examining racial disparities in drug arrests. *Justice Quarterly*, 32(2):288–313, 2015.
19. Jonathan Mummolo. Modern police tactics, police-citizen interactions, and the prospects for reform. *The Journal of Politics*, 80(1):1–15, 2018.
20. Emma Pierson, Sam Corbett-Davies, and Sharad Goel. Fast threshold tests for detecting discrimination. In *International Conference on Artificial Intelligence and Statistics*, pages 96–105, 2018.
21. Greg Ridgeway. Assessing the effect of race bias in post-traffic stop outcomes using propensity scores. *Journal of Quantitative Criminology*, 22(1):1–29, 2006.
22. Greg Ridgeway and John MacDonald. Doubly robust internal benchmarking and false discovery rates for detecting racial bias in police stops. *Journal of the American Statistical Association*, 104(486):661–668, 2009.
23. Jeff Rojek, Richard Rosenfeld, and Scott Decker. The influence of driver's race on traffic stops in missouri. *Police Quarterly*, 7(1):126–147, 2004.
24. Matt Ryan. Frisky business: race, gender and police activity during traffic stops. *European Journal of Law and Economics*, 41(1):65–83, 2016.
25. Camelia Simoiu, Sam Corbett-Davies, and Sharad Goel. The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics*, 11(3):1193–1216, 2017.
26. Michael Smith and Matthew Petrocelli. Racial profiling? A multivariate analysis of police traffic stop data. *Police Quarterly*, 4(1):4–27, 2001.
27. Rob Voigt, Nicholas P Camp, Vinodkumar Prabhakaran, William L Hamilton, Rebecca C Hetey, Camilla M Griffiths, David Jurgens, Dan Jurafsky, and Jennifer L Eberhardt. Language from police body camera footage shows racial disparities in officer respect. *Proceedings of the National Academy of Sciences*, 2017.
28. Patricia Warren, Donald Tomaskovic-Devey, William Smith, Matthew Zingraff, and Marcinda Mason. Driving while black: Bias processes and racial disparity in police stops. *Criminology*, 44(3):709–738, 2006.
29. David Word, Charles Coleman, Robert Nunziata, and Robert Kominski. Demographic aspects of surnames from census 2000, 2008. URL: http://www2.census.gov/topics/genealogy/2000surnames/surnames.pdf.
30. David Word and Colby Perkins. *Building a Spanish Surname List for the 1990's: A New Approach to an Old Problem*. Population Division, US Bureau of the Census Washington, DC, 1996.