Preliminary Draft

Measuring Racial Disparities in Police Stops: Evidence from Telemetric Mobility Data*

Xiaofeng Gong † Yuling Han ‡ Susan Parker § Matthew B. Ross ¶
Stephen L. Ross ‖

January 27, 2026

JEL Codes: K42, J15, C81, R41

Abstract

We address the longstanding methodological challenge of establishing a reliable counterfactual for evaluating racial disparities in traffic stops using telemetric data to estimate the racial composition of motorists. We develop and validate a dynamic measure of motorist racial composition using telemetric mobility data and a novel two-stage correction methodology anchored by an independent ground-truth measure, i.e. racial composition of motorists involved in accidents. Applying this to 2021–2022 Massachusetts State Police stops, we find Black and Hispanic motorists are stopped at rates exceeding their roadway presence by 4.4 and 4.9 percentage points, respectively. We demonstrate that uncalibrated telemetric data significantly understates minority presence (15.3 vs. 27.7 percent). Crucially, we leverage this validated measure to assess the accuracy of several commonly-used disparity tests. We find that the widely used "Community Standard" test is unreliable and dramatically overestimates disparities, even when restricting the sample to local roads or non-commuting hours. In contrast, the less-often used "Crash Benchmark" accurately captures the extent of disparities and remains robust even when aggregated to higher temporal levels covering up to 93 percent of stops. Finally, we show that the "Veil of Darkness" test yields smaller estimates than our telemetric measure, consistent with its focus on disparate treatment rather than the broader legal standard for measuring disparities in traffic stops. Our results suggest that telemetric data provides an effective metric for policymaking and litigation that complies with statutory mandates.

*Corresponding Author: Matthew B. Ross, Northeastern University, email: ma.ross@northeastern.edu
Acknowledgments: The authors wish to acknowledge Kenneth Barone, Associate Director of IMRP and Manager of CTRP3 at the University of Connecticut, for providing administration and funding to support this research project. All views expressed in this manuscript, and any associated factual errors, are our own.

†Gong: University of Connecticut xiaofeng.gong@uconn.edu
‡Han: Northeastern University han.yul@northeastern.edu
§Parker: Northwestern University susan.teresa.parker@gmail.com
¶M. Ross: Northeastern University ma.ross@northeastern.edu
‖S. Ross: University of Connecticut stephen.l.ross@uconn.edu

1 Introduction

The problem of racial disparities in policing constitutes a major challenge facing policymakers, with significant downstream social and economic impacts. Not only do these disparities generate deep economic costs (Mello, 2025), but they also erode public trust (Ang et al., 2021) and heighten the risk of severe harm to individuals (Tapp and Davis, 2024). Traffic stops represent the single largest point of contact between citizens and law enforcement, totaling over 21 million annual interactions (Pierson et al., 2020). Across the U.S., minority motorists have been consistently over-represented in traffic stops (Abrahams, 2020; Dunn, 2014; Pierson et al., 2020; Rojek et al., 2004; Smith and Petrocelli, 2001; Warren et al., 2006). In recent years, this over-representation has prompted a robust legislative response, including federal Title VI enforcement and state laws designed to monitor and prohibit profiling (e.g., Connecticut's Alvin Penn Act, Conn. Gen. Stat. 54-1 or California's Racial and Identity Profiling Act, Cal. Gov't Code § 12525.5).[1] State statutes often explicitly mandate that enforcement be evaluated relative to the underlying driving population.[2] While federal statutes are less explicit, Department of Justice enforcement has typically applied a similar standard.[3] Enforcing these mandates, however, remains challenging due to a longstanding methodological hurdle: the absence of a reliable measure of the demographic composition of drivers on the road.

The practice of comparing the demographic composition of traffic stops to the driving population is explicitly codified in relevant legislation and reflects a specific legal estimand we term "localized" disparate impact. This standard assesses whether minority motorists are stopped at disproportionate rates relative to their presence on the road at the specific times and locations where police are deployed. This concept is distinct from disparate treatment, which focuses on the differential treatment of similarly situated drivers based on race. Localized disparate impact is broader, capturing both intentional bias and the effects of race-neutral policies—such as the type and severity of violations triggering enforcement—that disproportionately affect minority motorists relative to their roadway presence. For instance, disparate treatment in speeding enforcement implies that officers stop minority motorists at lower speeds than their majority peers. In contrast, localized disparate impact applies to scenarios where a race-neutral directive—such as lowering the speed threshold for stopping *all* drivers—disproportionately affects minority motorists because their underlying speed distribution differs from that of majority drivers in a manner that increases their exposure. Finally, we include the qualifier "localized" to distinguish this metric from broader definitions of disparate impact that encompass upstream decisions regarding the location and timing of police deployment—factors often explicitly treated as conditional within the relevant legislation.

[1]There are 33 states with statutes addressing racial profiling, disparate treatment, or disparate impact in policing.

[2]Of the 33 states that have enacted statutes, 17 have explicitly mandated the statistical analysis or testing of stop data to identify racial disparities, racial profiling, or disparate impact. These jurisdictions include: California (Cal. Gov't Code § 12525.5), Connecticut (Conn. Gen. Stat. § 54-1m), the District of Columbia (D.C. Code § 5-113.01), Illinois (625 ILCS 5/11-212), Maine (Me. Rev. Stat. tit. 5, § 337-D), Maryland (Md. Code, Transp. § 25-113), Massachusetts (St. 2019, c. 122, § 10), Missouri (Mo. Rev. Stat. § 590.650), Nebraska (Neb. Rev. Stat. § 20-504), Nevada (Nev. Rev. Stat. § 289.820), North Carolina (N.C. Gen. Stat. § 143B-903), Oregon (Or. Rev. Stat. § 131.935), Rhode Island (R.I. Gen. Laws § 31-21.2-6), Texas (Tex. Code Crim. Proc. Art. 2.134), Vermont (Vt. Stat. Ann. tit. 20, § 2366), Virginia (Va. Code Ann. § 52-30.2), and Washington (Wash. Rev. Code § 43.101.410).

[3]For example, Department of Justice regulations implementing Title VI of the Civil Rights Act explicitly prohibit practices that have the 'effect' of discrimination, regardless of intent. Federal consent decree investigations and enforcement has also largely focused on addressing the disparate impacts of police enforcement policies rather than only overt disparate treatment or racial profiling (Walker, 2022).

Meeting the evidentiary threshold for localized disparate impact requires a precise counterfactual accounting for granular hour-by-hour and road-by-road variation in roadway demographics. In this paper, we address this challenge by developing a novel measure of the driving population using telemetric mobility data. Applied to Massachusetts State Police (MASP) stops from 2021–2022, our measure captures the vast majority of temporal and geographic variation in enforcement and documents a significant over-representation of Black and Hispanic motorists. We then employ this validated baseline as a high-precision 'ground truth' to evaluate several of the commonly used alternative tests for traffic stop disparities. While our telemetric approach provides a precise benchmark, its complexity limits its scalability for routine policy monitoring; therefore, a primary contribution of this work is determining which of the existing, cost-effective tests accurately reflect the relevant legal estimand. We find that the traditional Community Standard (Census) benchmark is largely unreliable, even when adjusted for commuting patterns. In contrast, the racial composition of parties involved in vehicle crashes provides an accurate measure of the racial composition of motorists on the roadway. While previously underutilized due to coverage concerns, we demonstrate that the crash benchmark maintains accuracy even when aggregated to higher temporal and geographic levels, allowing it to assess a large share of police stops. Finally, the "Veil of Darkness" (VOD) test yields smaller disparity estimates, consistent with its implicit focus on disparate treatment rather than the broader localized disparate impact required by the current legal standard.

Given the legal standards and practices, researchers and policymakers have historically faced a difficult choice: rely on the more rigorous subset of tests that capture only disparate treatment like VOD (Grogger and Ridgeway, 2006) and/or only prejudice-based discrimination like studies of search behavior and success (Anwar and Fang, 2006; Knowles et al., 2001), or revert to less scientific methods in an effort to align with the broader standard of the underlying driving population. When policymakers do opt for applying a broader standard of localized disparate impact, they are forced to rely on benchmarks that have historically been viewed with skepticism due to justifiable concerns regarding their accuracy or representativeness. For example, the "Community Standard" (i.e. Census-based benchmarks) approach is frequently used in litigation, yet its validity rests on the assumption that residential demographics predict on-road drivers (McDevitt 2001). Critics have rightly noted that this assumption is easily violated by racial differences in vehicle ownership, commuting and other travel patterns, and transit use, biasing tests for racial disparities (Grogger and Ridgeway, 2006). Alternatively, a smaller set of researchers have employed "Crash Benchmarks", utilizing not-at-fault collisions to approximate the driving population, as validated by Alpert et al. (2004) using direct observations of motorist race. While the assumptions underlying the Crash Benchmark are generally viewed as more appealing, crashes are relatively rare, historically limiting their spatio-temporal coverage for the monitoring of a broad population of police stops.

To address these deficiencies, the research community has recently begun to consider the use of high-frequency telemetric mobility data to construct a granular measure of the driving population.[4] This method offers a powerful exogenous measure of on-road presence that is independent of police interaction. Prior studies have established the potential of using telemetry to analyze speeding behavior and stop disparities; for example, Cai et al. (2022) utilized vehicle telemetry to compare the spatial distribution of speeding enforcement to the distribution of speeding behavior of motorists, while Aggarwal et al. (2025) and Xu et al. (2024) leveraged motorist location data from Lyft drivers and telemetric data, respectively, to analyze racial disparities in citations conditional on the racial composition of relevant motorists.[5] However, these foundational studies face significant

[4]Other data on travel patterns, such as LEHD Origin-Destination Employment Statistics (LODES) or the Census Transportation Planning Package (CTTP) only capture commuting trips between residential and workplace location, and only the CTPP separates trips by race and ethnicity.

[5]Other uses of telemetric data in the policing literature include Chen et al. (2025) who use telemetric

limitations regarding generalizability. The reliance on specific samples, such as ride-share drivers (Aggarwal et al., 2025), or restricted temporal windows, such as the use of predicted race for a typical Thursday in 2022 by Xu et al. (2024), limits the utility of these findings for broad, continuous policy monitoring. More generally, raw telemetric data, including the data utilized by Xu et al. (2024) and this study, suffers from two critical limitations: it does not explicitly identify race, requiring internal predictions of demographics, and it draws from an unknown selected sample of the population, e.g. telemetric data tends to under-represent low-income households and minority groups (Li et al., 2024; Jardel and Delamater, 2024).

Building on this important earlier work, we make three primary contributions. First, we enhance the prediction of motorist race by utilizing the racial composition of motorists specific to the time, date, and location of police stops, rather than relying on residential or workplace racial composition or on commuter composition for origin-destination pairs.[6] This allows us to better address concerns regarding selection into mobile device ownership or devices held by non-motorist vehicle occupants. Second, we expand the scope of analysis to include diverse conditions—such as rush hours, varying police shifts, and holidays—across the full police stop sample. Third, we validate our measure against traditional models that utilize high-dimensional fixed effects. Finally, we compare our mobility-based measure against three of the most commonly discussed approaches for assessing disparities in police stops: the Community Standard, the Crash Benchmark, and the Veil of Darkness (VOD).

For predicting the racial composition of motorists on the road, we introduce a novel two-stage methodology. First, we construct a baseline proxy for motorist racial composition by linking anonymized mobile device GPS pings to the racial and ethnic demographics of their inferred Census Block Group of residence, which is the most important internal source of information for predicting the race of mobile device users. We then create residential origin based proxies for the racial composition of motorists on each state route, highway, and interstate by town by date by hour of the day. We then predict the race of individuals involved in motor vehicle accidents using the residence-based racial composition of motorists on the road at the time of the accident. Unlike traffic stops, which involve significant officer discretion, crash involvement is largely determined by the event itself; thus, crash data serves as a low discretion 'ground truth' for who is actually on the road around the time that the stop was made.[7] To eliminate attenuation bias from measurement error driven by infrequent observations in the highly granular data, we employ a split-sample instrumental variables (IV) procedure, randomly dividing devices into analysis and holdout samples to eliminate attenuation bias from measurement error in our racial composition proxies. By regressing the race of crash-involved drivers onto our telemetric predictors using our two stage approach, we generate adjustment coefficients that statistically correct for prediction errors in motorist race that arise within many race-related applications of telemetric data.

To validate our measure of motorist racial composition, we first show that the measure can

data to study the deployment patterns and presence of patrol officers and Weisburd (2021) who leverages automatic vehicle locator (AVL) data GPS data from police patrol cars to examine the effects of police presence on crime.

[6] Athey et al. (2021) and Couture et al. (2025) rely on residential location identified within the telemetric data, while Xu et al. (2024) rely on vendor-provided predictions of race calibrated to the racial composition of telemetric-based residential and work locations, plus notably calibrated to commuting flows from the Census Transportation Planning Package (CTTP).

[7] As mentioned above, Alpert et al. (2004) uses observers to gather data on motorist race from eleven intersections in in Miami-Dade County. The percentage point differences between observed motorist racial composition and the racial composition of not-at-fault motorists involved in accidents was less than 2 percentage points, suggesting that accidents provide an accurate estimate of the racial composition of motorists.

explain variation in traffic patterns, explaining 54 to 65 percent of the variation in the racial composition of stops across towns, compared to residential town composition that only explains 2 to 20 percent. We also find that our measure explains 39-58 percent of the variation in the racial composition of stops across hour of the day and day of the week. Visually, the stop data indicates large variation in racial composition of stops across towns and over the hours of the day, and these patterns are closely matched by the telemetric based roadway composition. Next, we turn to the officer benchmarking approach of Ridgeway and MacDonald (2009, 2014). Specifically, in related work (Gong et al., 2025), we show that models including town by highway by shift fixed effects are sufficient to deliver estimates of officer propensities to stop minority motorists that are very robust to adding additional controls related to time of day, day of week and season of year.[8] In models without any controls, the average difference between the race of the motorist stopped and the predicted minority share on the road at the time of the stop yields a set of officers propensities that have a correlation ranging from 0.87 to 0.90 with those obtained from models that control for highway by town by shift fixed effects model. These exercises provide strong evidence that we have developed a powerful and valid benchmark for the driving population in a world without disparate impact or treatment which is capable of capturing granular spatio-temporal variation in exposure risk that is typically captured with high dimensional fixed effects.

Applying this externally motorist composition calibrated, telemetric measure to MASP traffic stop data delivers four key empirical findings. First, we confirm clear evidence of localized disparate impact: minority motorists are stopped more frequently than their presence on the road predicts. Specifically, we find that Black and Hispanic motorists are stopped at rates exceeding their presence on the roadway by 4.4 and 4.9 percentage points, respectively. Our evidence demonstrates that the telemetric benchmark accurately captures the complex fluctuations of police activity, closely mirroring the 'shape' of enforcement throughout the week, across the state and across officers, and yet we document a persistent 'level shift' where the volume of stops for Black and Hispanic motorists consistently exceeds the predicted non-discriminatory baseline. We find that external calibration to motorist roadway composition is critical; estimates based solely on telemetric data suggest minority motorists comprise only 15.27 percent of the driving population, compared to 27.71 percent with the calibrated measure. The composition prior to the accident based calibration implies a minority disparity of 19.5 percentage points, similar to the 15 to 20 percentage point Black stop disparity in Xu et al. (2024), while our calibrated measure implies a disparity of only 7.1 percentage points.[9] Notably, our telemetric data implies a substantially higher share of Black and Hispanic drivers at night, which cannot be captured by data on the racial composition of commuters as utilized for the vendor provided data used by Xu et al. (2024).

Second, while the Community Standard Benchmark may point to patterns in the location and timing of stops that should be investigated, our analyses show that the widely used Community Standard dramatically overestimates the magnitude of the disparity by 7 to 13 percentage points when the legal standard is the composition of motorists at the time and location of the stop. By relying on residential rather than on-road demographics, the Census approach fails to account for the large variation in the racial composition of motorists over space and time, leading to inflated estimates of stop disparities and, consequently, systematically erroneous policy conclusions about the extent of the disparity, as measured based on current legal standards. While the community standard has been frequently criticized, actual assessments of the standard are quite scarce. Further,

[8]Additional vectors of fixed effects based on this information yield officer propensities that have correlations of 0.98 or higher with the propensities arising from the simple two-way fixed effects model that only includes officer and town by highway by shift fixed effects.

[9]Our estimates are moderately larger than the disparate treatment estimates of discrimination in speeding stops against minority Lyft drivers of 3 to 4 percentage points (Aggarwal et al., 2025).

lower car ownership and higher transit use among Black households would tend to suggest that a community standard would overstate the share of Black motorists on the road (Bunten et al., 2024). Key concerns such as commuters and other motorists traveling through a community might potentially be mitigated by, for example, excluding stops made during rush hour or on major interstates. However, in our analysis, these types of adjustments to the sample did nothing to change the empirical reality that community racial composition substantially understates the share of minority motorists in the locations and times when police stops tend to be made.

Third, we establish that the under-utilized Crash Benchmark accurately captures the extent of localized disparate impact and can be successfully extrapolated to provide relatively comprehensive coverage of police stops. While critics often cite the rarity of crashes as a limitation, we find that crash data need not be matched at a hyper-granular level to be effective. Even when aggregated to the highway by town by shift level, which expands coverage from 2 percent to 93 percent of stops, the crash benchmark generates disparity estimates that are generally consistent with our granular (highway by town by date by hour) telemetric measure. The crash benchmark estimates remain close in magnitude to our telemetric benchmark estimates as we increase temporal aggregation to expand the share of stops covered, even though the disparities estimated for Black motorists increases as we expand the sample. Note that we include all motorists involved in accidents in our analysis, rather than not-at-fault accidents, given concerns that the assignment of 'not-at-fault' is potentially endogenous to officer reporting bias (West, 2018).

Finally, the Veil of Darkness (VOD) test consistently estimates a smaller disparity, reinforcing the distinction between measuring racial bias (i.e. disparate treatment) and broader impacts of overall enforcement (i.e. localized disparate impact). Our analysis suggests one possible explanation for this divergence: police response to visibility; for instance, speeding enforcement increases during daylight hours across all demographic groups relative to other moving violations, complicating the VOD's ability to isolate disparate treatment from changes in enforcement behavior. Nonetheless, the VOD estimates of disparate treatment consistently understate the target policy parameter which is the total disparate impact experienced by minority motorists relative to their actual presence on the roadway when and where enforcement is occurring.

These findings carry immediate policy implications. Our measure provides a superior metric for public policy and litigation, fulfilling legal standards more reliably than previous comparisons. Designed to be computationally efficient, our application of telemetric data when combined with some external benchmark for race for limited locations and/or times, such as accident data or perhaps traffic camera data, can be directly incorporated into police monitoring systems. In addition, the disaggregated nature of the telemetric-based measure also avoids the need for high dimensional controls that often become problematic when evaluating the stop patterns of small departments, individual officers and specific locations where the number of stops can become quite thin. However, our analyses also strongly suggest that the Accident or Crash Benchmark should be actively considered as an additional, potentially more cost-effective strategy for government monitoring and evaluation of racial disparities in police stops.

The remainder of the paper is organized as follows. Section II describes the data, comprising the telemetric mobility data, MASP traffic stop records, and statewide crash records. Section III details our novel two-stage methodological strategy for constructing the exposure rate counterfactual. Section IV applies this counterfactual to MASP enforcement data to estimate the degree of localized disparate impact and validates the reliability of our measure. Section V leverages this validated benchmark to evaluate the accuracy of alternative metrics commonly cited in the literature. Finally, Section VI concludes with a broader discussion of policy implications.

2 Data and Sample Construction

We combine data from several sources from 2021-2022 to develop and validate our telemetric benchmark, to document racial disparities in traffic stops in Massachusetts, and to compare our telemetric benchmark to traditional benchmarking approaches, i.e. community standards based on town residential composition, racial composition of accidents, and the Veil of Darkness (VOD) approach that exploits reduced visibility of race at night.

2.1 Massachusetts Traffic Stop Data

Police stop data in Massachusetts were obtained from several law enforcement sources to form a sample of police traffic stop encounters that includes written traffic violation citations, written warnings, and verbal warnings from 2021-2022. Written violation data were obtained from Massachusetts State Police (MASP) division troops A, B, C, D, and H. We also obtained a statewide database of timestamped state police interactions. Because written violations did not have a precise hour timestamp, we matched written violations to a timestamped state police interactions database to obtain the time of the event. We restrict the sample to written citations and warnings that were successfully merged between the two data sets, about 416,380 stops or 87.4 percent of the total.

Separately, we received a data set containing verbal warnings from Massachusetts State Police. The merged written violation stops and the independent verbal warning stop samples both contain the officer who made the stop, the precise date and time of the stop, the highway on which the stop was made, the town or city in which the stop was made, as well as the race and gender of the motorist. However, unlike for written violation stops, the verbal stop data does not identify the reason for the stop.

Finally, we downloaded geographic roadway data from the Massachusetts Department of Transportation (MASSDOT), and restrict our sample of stops to those occurring on any roadway that MASSDOT identified as an Interstate, U.S. Highway or State Route. We also drop any stops that were identified as occurring on residential surface streets. These restrictions eliminated about 19.7 percent of our combined stop sample.

We collapse this data into cells that identify the location of the stop based on the specific highway and the town in which the stop was made and the time of the stop identified as occurring during a specific hour of the day on a specific date. For these cells, we calculate 1. the fraction of stops that are of non-Hispanic Black motorists among stops of non-Hispanic White and Black motorists, 2. the fraction of stops that are of Hispanic motorists among stops of non-Hispanic White and Hispanic motorists, and 3. the fraction of stops that are of non-Hispanic White motorists among Hispanic motorists and non-Hispanic White and Black motorists, as well as the number of stops of motorists each of Hispanic and non-Hispanic White and Black motorists and the share of stops that are of female motorists. Many cells contain a single stop leading to fractions of zero or one, and the fraction is missing if the cell does not contain any stops of motorists of the races used to form the denominator of a given racial share.

2.2 Massachusetts Traffic Crash Data

We additionally obtain motor vehicle crash data from Massachusetts. The crash data also contains the date, time, specific roadway, and the town of the accident as well as driver race for about half of the accidents. We expand the crash data to contain one observation per motorist involved in a given accident. We restrict the crash data to accidents that occurred between 2021-2022 and to accidents where motorist race was recorded by the responding officer, about 46 percent of the

accidents. Finally, as with the stop data, we restrict the sample to accidents recorded as occurring on an Interstate, U.S. Highway or State Route, dropping accidents where the officer identified the accident as occurring on an on or off ramp.

2.3 Telemetric Data

We use data from a geospatial data firm that provides access to anonymized and privacy-preserving telemetric data. Telemetric data consists of longitudinal geospatial data from mobile phones, including both IOS and Android devices. Mobile phones are uniquely identified using a privacy-preserved identifier which is a hashed and anonymized ID of each device's mobile advertising identifier (MAID) along with corresponding GPS "pings", or a timestamped geographic location (latitude and longitude). All users provide informed consent to anonymous data collection for research. To further preserve privacy, the vendor suppresses precise home locations of devices in its panel, and provides the Census block group where the individual using the digital device is expected to reside, and in many cases the Census block group where the individual is expected to work. These residential and work locations are based on the daily pattern of observed "pings". Additionally, location data associated with sensitive locations such as schools, religious institutions, certain medical locations, and other privacy-sensitive areas is removed from the dataset by the data vendor. intervals. The data vendor contractually prohibits researchers from attempting to re-identify users, or to probabilistically infer direct 1:1 device matching between the telemetric data and other datasets. Other than the location and pattern of pings, we do not observe any information about the individual using the mobile device.

We construct a sample of mobile devices that we observe consistently residing in Massachusetts and in immediate bordering counties that may reasonably commute to/from or through Massachusetts from 2021-2022. We include mobile devices that we observe residing within Massachusetts and the bordering counties. We restrict the sample to mobile devices we observe residing in Massachusetts, Connecticut, Rhode Island, New York, Vermont or New Hampshire. We retain only those mobile devices that reside within the state or bordering counties for the consecutive three-month span. For instance, a mobile device we observe in Worcester, Massachusetts for 2 consecutive months and Tampa, Florida for the 3rd month would be excluded from our sample. The geospatial firm providing data re-assigns mobile devices to a new internal identifier if the device remains in the sample for longer than 365 days, and so the same individual may appear multiple times under different device identifiers. [10]

For each mobile device that remains consistently in our sample, we geocode all associated GPS pings to the roadway network. As discussed above, the roadway network is based on all state routes, interstates, and U.S. interstates identified in the MASSDOT geographic information files. We drop segments corresponding to ramps and tunnels. GPS pings contain granular information on a device's latitude, longitude, speed, and direction recorded at rapid time intervals. To prepare these roadways for spatial matching to GPS pings, we convert each roadway segment, originally a one-dimensional linestring, into a polygon by adding a 32-foot buffer on either side.[11] This buffered roadway polygon approximates the drivable width of major highways and allows us to identify whether a given GPS ping falls within the roadway. We then assign each GPS ping from consistent devices to the corresponding buffered roadway polygon.

[10] We observe a small proportion of our sample that remained in the sample for longer than a year. We omit these anomalous observations.

[11] We visually inspected the addition of the 32 foot buffer in several large highway settings and found the width was able to accommodate the highway expanse without including parallel service roads or other nearby roadways.

To validate that our matched pings reflect actual roadway travel, we examined the empirical distribution of the distance between the road linestring and an observed ping located within buffered highway polygons. While the data is thinner than our full telemetric sample, some pings identify an instantaneous travel speed and we limit our analysis of highway width to pings with an observed speed that were traveling at least 60 miles per hour, or a speed unlikely to occur off the highway. Pings traveling at least 60 miles per hour on average fell within approximately 32 feet of either side of the roadway line string, suggesting that the buffered polygon captures a high proportion of true highway GPS pings.

Unfortunately, we are not able to distinguish if the mobile device is from a passenger in a vehicle or a driver, or if the mobile device is on public transportation such as a bus or other taxi or ride-sharing transit. This limitation is one of the key reasons behind the need to calibrate the telemetric data to an outside measure of motorist racial composition. We also note that public transportation ridership is a small fraction of each state's population in most parts of Massachusetts with the exception of the Boston area (Massachusetts Department of Transportation, 2012).

As with the stop data, we collapse the telemetric data into cells that identify the location of the stop based on the specific highway and the town in which the stop was made and the time of the stop identified as occurring during a specific hour of the day on a specific date. For each cell, we calculate proxy variables for the roadway composition as described below.

2.4 Telemetric-based Roadway Composition

In the vast majority of studies using telemetric data, including ours, the race of the mobile device user is unobserved. We utilize racial composition of the residential Census block group to characterize the distribution of motorists on a given roadway segment and time since residential location is by far the strongest predictor of race in telemetric data.[12] Following the structure of the stop and crash data, we create our measure of roadway composition by aggregating telemetric data from unique mobile devices observed on the same roadway and in the same Massachusetts town for a specific hour, day, month and year of our sample, a unit of analysis we refer to as a cell.

In each cell c, our goal is to predict the fraction of motorists who belong to a racial group as a share of that group and the comparison racial group, i.e. non-Hispanic White for non-Hispanic Black and Hispanic motorists and Black and Hispanic for non-Hispanic White motorists. As a starting point, we assign an initial race likelihood L_{ij} to each device id i based on the share of total population in that Census block group that belongs to race j. Using these values, we create a proxy for the number of motorists on the roadway associated with that race N_j by summing these likelihoods over all device id's.

$$N_j = \sum_{i \in c} L_{ij} \tag{1}$$

For each cell, our proxy for the fraction of the relevant motorist racial group Pr_{cj} is simply

[12]The telemetric data utilized by Xu et al. (2024) contains a prediction of the racial share of motorists on a roadway that utilizes both work place and residential location. However, using LODES data, we observe that work place racial segregation is quite modest relative to the high levels of residential segregation observed, at least in our sample of states. Notably, Xu et al. (2024) also utilizes traffic camera data assigning race based on the racial composition of the residential location to which ticket was mailed.

$$Pr_{cj} = \frac{N_j}{N_j + \sum_{k \in \Omega} N_k} \tag{2}$$

where Ω is the set of comparison racial groups, i.e. non-Hispanic Whites for Hispanic and non-Hispanic Black stop outcomes and Hispanic and non-Hispanic Blacks for non-Hispanic White stop outcomes. As noted previously and discussed subsequently in the empirical section, we will adjust these home-location-based proxies based on the observed composition of the actual race of motorists involved in accidents within these cells.

2.5 Analytic Samples

We create two distinct analytical samples, a crash sample and a stop sample. For the crash sample, we split the telemetric data by device id into an analysis and a hold-out sample. For each motorist in the crash sample, we merge in two roadway composition variables created using the analysis and the hold-out sample by cell, i.e. highway by town by hour by date. This data will be used to estimate models of the race of the motorist involved in an accident using split sample instrumental variables analysis to avoid attenuation bias from measurement error. The crash data are further organized into three overlapping subsamples containing 1. Non-Hispanic Black and White motorists, 2. Hispanic and non-Hispanic White motorists, and 3. Non-Hispanic White and Black motorists and Hispanic motorists.

For the collapsed stop sample, we merge the cell roadway composition variables based on the entire telemetric data sample to obtain the most accurate measure of device id's observed for each cell. We then use the estimates from the crash data based race predictive models to adjust the telemetric based roadway composition proxies. As with the crash sample, the cell based stop data will be restricted to cells that 1. contain at least one non-Hispanic White or Black stop for analyzing disparities faced by Black motorists, 2. contain at least one Hispanic or non-Hispanic White stop, and 3. contain at least one non-Hispanic White or Black or Hispanic stop.

3 Empirical Strategy

We start by using the racial composition of motorists involved in accidents to calibrate the residential location based proxies for the racial composition of motorists. This calibration relies on three assumptions: 1. accidents are randomly assigned to motorists, 2. the devices observed in the telemetric data provide a representative sample of motorists on the road, and 3. cell level deviations of the motorist population composition from the residential composition are random. While we cannot test these assumptions directly, we will provide evidence below that the calibrated racial composition captures a substantial share of the variation that is often captured by high dimensional fixed effects in models of police stops. Further, given our use of the crash data to calibrate the racial composition proxies, we do not require that assumptions 2 and 3 hold in terms of levels. On average minority motorists could be systematically underrepresented in the telemetric data or could be overrepresented by population estimates due to, for example, low rates of car ownership or high rates of mass transit usage.

However, both the imperfect coverage of the telemetric data and random variation in the share and travel patterns of drivers in different residential locations, as well as our hour long time window for the telemetric data, could all represent substantial sources of measurement error even if these assumptions hold. Therefore, we will use a split sample IV approach to estimate the relationship

between the race of each motorist m involved in an accident within cell c and the racial composition proxy Pr_{cj}.

The first stage model regresses the composition proxy from the hold-out sample Pr^H_{cj} on the proxy from the analysis sample Pr^A_{cj}. The model is estimated using the sample of motorists involved in accidents and conditions on a vector of high dimensional fixed effects d intended to capture broader patterns of police exposure to minority motorists. Specifically,

$$Pr^H_{mcdj} = \alpha_j Pr^A_{cdj} + \delta_{dj} + \mu_{mcdj} \tag{3}$$

where fixed effect clusters d are defined as highway by town by officer shift, i.e. day, evening and night, which has been shown to capture most of the systematic variation in the racial composition of stops and also does not split stops across cells. Gong et al. (2025) show that this vector of fixed effects produces robust estimates of officer minority motorist stop propensities that are very stable to the inclusion of additional controls, e.g. correlations of officer propensities across models exceeding 0.98. Finally, note that while Pr^H_{mcdj} does not vary across motorists, the subscript m is included to represent a regression using the sample of motorists involved in accidents and the proxy matched to that motorist from the hold-out sample.

We next use the individual race of motorists involved in an accident R_{mcdj} as an outcome where the variable takes the value one if the motorist is of race j and zero if the motorist belongs to the comparison race or races. We use the predicted telemetric roadway composition fitted in the prior regression as an independent variable. For each race/ethnicity subsample j from the crash data, we separately estimate

$$R_{mcdj} = \beta_j(\hat{\alpha_j} Pr^A_{cdj} + \hat{\delta}_{dj}) + \gamma_{dj} + \epsilon_{mcdj} \tag{4}$$

where the parameter β scales the composition proxy to match observed racial composition shares on the roadway. The inclusion of town by highway by shift FE's assures that the calibration is identified by higher frequency variation within location and time frames, and is not simply driven by cross-sectional variation across the state or, for example, between accidents during rush hour and late at night. We cluster standard errors at the level of the fixed effects d.

Equation (4) can be estimated two ways by either following the equation exactly or by reversing the dependent and independent variables and regressing the proxy from the analysis sample on the proxy from the hold-out sample. With two sets of estimates in hand for β and the fixed effects γ_d, we estimate the racial composition for each cell and racial group as a weighted average of those estimates using the composition proxy from the full telemetric sample Pr^F_{cdj}.

$$\hat{R}_{cdj} = (\beta^H_j \omega_j + \beta^A_j (1 - \omega_j)) Pr^F_{cdj} + (\hat{\gamma}^H_{dj} \phi_{dj} + \hat{\gamma}^A_{dj} (1 - \phi_{dj})) \tag{5}$$

where the weights are based on the relative inverse variance of the estimates so that

$$\omega_j = \frac{1}{1 + \sigma^2_{\beta^H_j} / \sigma^2_{\beta^A_j}} \tag{6}$$

and

$$\phi_{dj} = \frac{1}{1 + \sigma^2_{\gamma^H_{dj}} / \sigma^2_{\gamma^A_{dj}}} \tag{7}$$

Rather than including high dimensional fixed effects, as is the standard in police stop models, we simply subtract the predicted racial composition within each cell from the racial composition of stops observed within each cell S_{cdj}. The resulting regressions using these differences as a dependent variable test the null hypothesis for whether police stop behavior deviates from the racial composition of motorists under the maintained assumptions described above.

$$(S_{cdj} - \hat{R}_{cdj}) = \theta_j + \xi_{cdj} \tag{8}$$

where the estimate of the intercept θ_j captures racial disparities in stops that are not explained by the racial composition of motorists. The regression is weighted by the number of stops in each cell following Donald and Lang (2007), and standard errors are clustered at the level of the fixed effect vector utilized in the calibration IV models above to allow for correlation in stops over space within shifts.

4 Main Results

4.1 Estimates of Racial Composition

In Table 1 we present first stage estimates from regressing the proxy for highway by town by date by hour racial composition from the holdout sample on the equivalent proxy from the analysis sample. The telemetric data was divided into two subsamples based on device id, and the first column for each racial group analyzed represents using sample 1 as the hold-out sample, while the second column uses sample 2 as the hold-out sample. The first stage is very powerful with F-statistics ranging between 174 and 307, but the IV estimates range between 0.23 and 0.32 implying substantial measurement error in the racial composition proxies. Therefore, a simple OLS regression of accident race on the proxy would suffer from substantial attenuation bias.

In Table 2, we present IV estimates from regressing the race of motorists involved in traffic accidents on the racial composition proxy where again the first column uses sample 1 as the endogenous right hand side variable and sample 2 as the instrument. In the bottom row of the table, we present the stop weighted mean of fixed effect estimates for each regression. In both columns, the coefficients are similar in magnitude and statistically significant for each sample, and the instrument has reasonable power, with F-statistics ranging between 20 and 60. Notably, the estimates for a Black motorist among Black and White motorists involved in accidents are substantially larger than the estimates for White motorists suggesting that a failure to externally calibrate telemetric proxies may lead to a substantial understatement of the exposure of police officers to Black motorists at the time and location of police stops, at least in our Massachusetts sample. On the other hand, the Hispanic estimates are closer to one, and the under-representation of Hispanic motorists in the telemetric data is captured by fixed effect estimates that are positive on average. Similarly, the White slope estimates relative to Black and Hispanic motorists are 1.35 and 1.61, and the White under-representation in the telemetric data is captured entirely by fixed effect estimates with large negative means of -0.42 and -0.64, respectively.

4.2 Validation of Racial Composition

We first demonstrate that estimates of racial composition have substantial power to explain a portion of the racial variation in police stops, given that police stops should be heavily driven by motorist composition whether or not disparities exist. We collapse the combined stop and telemetric data to three different levels: 1. town, 2. hour by day of week, and 3. week by year. We then calculate cell-level means weighted by the number of stops. These measures capture the fraction of stops and the predicted fraction of motorists in each racial group. Racial groups are defined as: the fraction non-Hispanic Black as a share of non-Hispanic Blacks and Whites, the fraction Hispanic as share of Hispanics and non-Hispanic Whites, and non-Hispanic Whites as a share of Hispanics and non-Hispanic Whites and Blacks.

Table 3 presents the standard deviations of both the racial composition of stops and the deviation between the racial composition of stops and the composition of motorists based on telemetric data where standard deviations are weighted based on the number of stops associated with each of these aggregated cells. Panel 1 presents the standard deviation across towns, panel 2 over the hours of the day and days of the week, and panel 3 over weeks throughout the two years of our sample. The town panel also presents the standard deviation for the share of stops minus the community racial share for the town in which the stop was made. The columns present the standard deviations for the non-Hispanic Black, Hispanic and non-Hispanic White shares. For all three racial groups, the control for the estimated racial composition of motorists substantially reduces the standard deviation across towns by between 54 and 65 percent. In contrast, adjusting for the town racial composition reduces the across town variation in stop racial composition by much less, by 20 percent for Black, 8 percent for Hispanic and only 2 percent for Whites. Similarly, adjusting for motorist composition reduces the variation across hours of the day and day of week by between 39 and 58 percent. The estimated roadway composition explains less of the variation over time using the sample by week by year with reductions in the standard deviation ranging between 18 and 36 percent.

We next present the information from Table 3 in graphical format. Figures 1-3 present the racial composition of highway stops in each town, the racial composition of the town residents, and the racial composition of motorists on highways in these towns for non-Hispanic Black, Hispanic and non-Hispanic White motorists, respectively. Dark green represents a higher share of the subject racial group, while dark blue represents a lower share.

Figures 1-3 show that the town racial composition is not only less correlated with racial composition of stops, but also understates the share of minorities observed on roadways where and when stops are made. Overall, panel 1 for stops and panel 3 telemetric motorist composition in Figures 1 and 2 show noticeably more light blue and green associated with more minorities than Panel 2 where the stop composition net of residential composition is presented. Similarly, Panels 1 and 3 of Figure 3 for Whites show more lighter green and more blue consistent with more Whites. Further, Panels 1 and 3 tend to depart from Panel 2 in the same regions or set of towns within the state.

Figure 4 presents the racial composition of both highway stops and motorists at the time and day of highway stops starting between midnight and 1:00 AM on Monday morning and running until between 11:00 PM and midnight on Sunday evening. Panels 1 through 3 present results separately for Black share, Hispanic share and White share. The minority share varies dramatically over each 24 hour day peaking between 2:00 and 3:00 AM and falling to much lower levels for the morning rush hour and then rising again after the evening rush hour. The time pattern of the racial composition of motorists closely mirrors the time pattern of the stop composition, while the somewhat lower minority shares for the motorist composition is consistent with racial disparities in stops.

Figure 5 presents the week by week variation in racial composition starting from the first week

of January through the last week of December plotting separate lines for the racial share in 2021 and 2022. By week, the racial composition of motorists lies substantially below the composition of stops, around 5 percentage points, and while the week to week fluctuations appear similar the racial composition of stops exhibits substantially more variation across weeks. This variation could arise from changes in policing as police departments often respond to events or federal funds by initiating short run, focused enforcement campaigns, like for seat belt violations, use of cell phones or driving while intoxicated.

Having demonstrated that our racial composition estimate captures substantial variation across both towns and time of day, we next assess whether the measure is able to successfully isolate or partial out variation in racial composition that appears unrelated to officer behavior. To do this, we draw on the officer benchmarking approach of Ridgeway and MacDonald (2009, 2014) and estimate models of officer propensity to differentially stop minority motorists relative to other officers.[13] In our own work (Gong et al., 2025), we have shown that controlling for highway by town by shift fixed effects are sufficient to generate estimates of officer propensities to stop minority motorists that are very stable to additional controls. Specifically, adding fixed effects for day of week by time of day, e.g. identifying morning and evening rush hour, and for month by year to capture both seasonality and trends over time generates officer propensities that correlate at levels of 0.98 or higher with the propensities conditional only on the highway by town by shift fixed effects.

We use the statewide cell level stop composition samples to estimate two models using as outcomes both the racial composition of stops in the sample of cells and the stop racial composition net of the estimated motorist composition. The first model, or model a, simply controls for officer fixed effects as indexed by o.

$$y_{ocdj} = \theta_{oj} + \eta_{ocdj} \tag{9}$$

Model b also includes highway by town by shift fixed effects.

$$y_{ocdj} = \theta_{oj} + \lambda_{dj} + \eta_{ocdj} \tag{10}$$

As above, these models are estimated weighting the samples by number of stops per cell. Then, following Weisburst (2022) and Gong et al. (2025), we Bayesian shrink these estimates to address the high levels of estimation error that can arise for officers with relatively small numbers of stops.

The officer fixed effects estimates are plotted in Figure 6 with panels a, c and e presenting the estimates the racial composition of stops for Black relative to White motorists, Hispanic relative to White motorists and White relative to Black and Hispanic motorists, and panel b, d, and f presenting the stop composition net of the motorist composition. Model a is presented, in blue and model b in yellow for each panel. The fixed effect means and standard deviations (in parentheses) are shown in the upper right hand corner of each panel. For all three samples in panels a, c and e, the inclusion of the highway by town by shift fixed effects dramatically changes the distribution of officer propensities for stops be of minority motorists from visibly highly skewed distributions with sizable standard deviations ranging between 0.105 and 0.134 to a relatively symmetric distribution with notably less variation, reductions between 49 and 57 percent. On the other hand, the inclusion of the incidental fixed effects to the model for stop composition net of motorist composition at the precise date and hour of the stop has minimal visible impact on the distribution of estimated,

[13]Also see Feigenberg and Miller (2022) and Goncalves and Mello (2021) for other examples of across officer assessments of racial disparities.

shrunken FE's and leads to much smaller reductions in the standard deviations of officer propensities, ranging between 12 and 18 percent. More importantly, the estimated officer stop propensities from model b are strongly correlated (ranging between 0.87 and 0.90) with the unconditional net of motorist composition officer stop propensities from model a. These high correlations suggest that the motorist composition captures most of the variation associated with the high dimensional fixed effects, and does so without overfitting concerns that typically arise with controls for high dimensional fixed effects.

4.3 Statewide Estimates of Racial Disparities

Again using the statewide highway stop sample and weighting by number of stops per cell, we regress the racial composition of stops net of the predicted racial composition of motorists for the cell to which each stop belongs on a constant term with no other controls. The likelihood of a stop involving a Black motorist (relative to Non-Hispanic White) is 4.4 percentage points higher than the predicted roadway composition, relative to a base black share of 19.88 percent in the stop data. The likelihood of a stop involving a Hispanic motorist (relative to Non-Hispanic White) is 4.9 percentage points higher than the predicted roadway composition, relative to a base share of 22.14 percent in the stop data. Finally, the likelihood of a stop involving a non-Hispanic White motorist (relative to Hispanic and Black) is 7.1 percentage points lower, relative to a base share of 65.25 percent in the stop data. All three are highly statistically significant with standard errors clustered at the highway by town by shift level.

In comparison, Xu et al. (2024) estimates of disparities are substantially larger. They plot the fraction of Black stops against the predicted share of Black motorists in their Figure 3, and over most of the distribution of motorist racial composition stop share Black exceeds the share of motorists by between 15 and 20 percentage points. Notably, our uncalibrated telemetric data predicts an average minority motorist share of 15.3 percent relative to the minority stop share of 34.8 percent, which would imply a similar 19.5 percentage point disparity. On the other hand, the calibrated share of minority motorists is 27.7 percent, which implies a 7.1 point disparity identical to the -7.1 point estimates for white disparities above. While our estimates of stop disparities are for state police in Massachusetts and Xu et al. (2024) study police stops in Chicago, the magnitude differences in the disparity estimates are unusually large.

There are several other alternative explanations for these diverging disparity estimates. Although the race prediction algorithms in the Replica data used by Xu et al. (2024) are proprietary, two key features are known: (1) they rely on the Census Transportation Planning Package (CTPP) which are generated from the residential and workplace location in the American Community survey as the only origin-destination based information on motorist racial composition; and (2) these predictions are aggregated to represent a "typical Thursday". As we demonstrated earlier, telemetric data reveal substantial temporal heterogeneity in motorist and stop composition—both diurnally and seasonally—that is masked by the static predictions in the Replica data. As a result, their estimates may be upwardly biased if the share of Black motorists increases at night relative to morning and evening commutes (as in our sample). Alternatively, their results may capture a broader definition of disparate impact driven by the macro-allocation of police resources,[14] rather than the "localized disparate impact" standard defined earlier.

In addition, the CTPP origin to destination commuting data is based on the American Community Survey question on where an individual worked last week, which leads to several limitations: 1. a substantial share of workers who were regularly employed last year were not employed last

[14]See Bushway et al. (2025) for a broader discussion of disparate impact in criminal justice.

week, 2. a large fraction of workers report either working at home or no fixed work place last week, even if the individual may have had a fixed work place other weeks during the year, and 3. this question has a relatively high rate of non-response leading to reliance on hot deck imputations. All these limitations may affect the composition of commuters represented in the CTPP. Such selection might reasonably be expected to correlate with job stability or with employment in professional occupations or office environment and so might affect the estimated racial composition of commuters.[15]

A final potential explanation is that our measure might underestimate racial disparities if minority motorists are overrepresented in crash data. However, the empirical literature contradicts this possibility. In particular, Alpert et al. (2004) finds that the divergence between observed motorist racial composition and that of not-at-fault motorists in accidents is less than 2 percentage points. Moreover, evidence on driving behavior suggests that minority motorists do not exhibit riskier driving patterns: Cai et al. (2022) show that speeding rates are comparable across predominantly White and minority neighborhoods, while Kalinowski et al. (2025) find that minority motorists actually drive slower and are involved in fewer fatal accidents during daylight—consistent with a behavioral response to perceived risk of racial profiling when race is visible.

5 Comparison to Alternative Benchmarks

We now repeat the analysis of racial disparities above tailoring the sample to match the sample used for each of the alternative benchmarks. Using these samples, we again regress the racial composition of stops net of motorist composition on a constant and cluster standard errors by highway by town by shift, except for our last analysis where we change the clustering based on the Veil of Darkness structure of the sample.

5.1 Community Standard

For the community standard, we append the racial composition based on the town in which the stop was made using aggregate data from the American Community Survey 2018 five year moving average. We then regress the racial composition of stops in a cell minus the telemetric-based roadway composition or the town racial composition on a constant term. Further, to address concerns that the town racial composition may understate the share of minorities on the roadway due to commuters or other through traffic, we restrict the sample for both the telemetric and the town benchmark regressions by removing stops made during the morning and evening rush hour, and then removing stops made on interstate highways.

These results are presented in Figure 7 with panel a presenting estimates for non-Hispanic Blacks relative to non-Hispanic Whites, panel b for Hispanics relative to non-Hispanic Whites and panel c presenting non-Hispanic Whites relative to Hispanic and non-Hispanic Black motorists. For the full sample, the community standard based estimates are 14.2 percent relative to telemetric-based estimates of 4.4 percent for Blacks and 11.9 percent relative to a telemetric estimate of 5.5 percent for Hispanics. Dropping rush hour stops and dropping stops on the interstate does little to alter this clear pattern. Dropping rush hour stops increases disparities for both the telemetric and community standard disparities and dropping major interstates decreases disparities, but the gap between the telemetric and community standard measures are relatively unchanged. Therefore, in spite of on average lower rates of car ownership and higher usage of mass transit among Black

[15]See Green et al. (2017) who compare the CTPP to commuting flow data based on state Unemployment Insurance records from the LEHD Origin-Destination Employment Statistics (LODES).

and Hispanic motorists, the community standard appears to consistently overstate racial disparities compared to measures based on the composition of motorists on the roadway at the time of the police stops.

5.2 Crash or Accident Benchmark

We next examine the crash or accident benchmark directly. In our calibration exercise, we used crash data to estimate a model to adjust the residential location based estimates of roadway composition calculated from the telemetric data. With a predictive model in hand, we are able to apply it to the rich telemetric data for virtually all highway by town by date by hour cells in which police stops are observed. In contrast to this approach, the accident benchmark has typically been applied to only a small share of police stops made at the location and the general time frame when accidents occurred in order to capture information about the composition of the roadway during the accident. We start with a very similar exercise estimating the models for the racial composition of stops both net of the racial composition of motorists involved in accidents and net of the telemetric roadway composition estimate for a sample of highway by town by date by hour cells in which accidents and police stops both occurred. Given our use of the accident data to calibrate the model, it would not be surprising if the estimates of racial disparities are similar for the accident and telemetric benchmark, and as shown by the top estimates in each panel of Figure 8 the estimates are quite similar. Specifically, Black disparities are 3.5 percent using the accident benchmark and 2.9 percent using telemetric data, and Hispanic disparities of 4.9 percent using accident and 6.0 using telemetric data, which constitute differences that are easily within the standard errors of the disparity estimates.

However, the critical question is whether the estimates remain similar as we expand the sample time frame to allow for more stops to be covered by accidents, while continuing to use our telemetric benchmark based on roadway composition at the highway by town by date by hour. Specifically, for the second sample and second estimates from the top of the figure, we create larger hour bins defined as 2am-5am, 6am-10am, 11am-3pm, 4pm-8pm and 9pm-1am to seperate morning and evening rush hour from daytime and nightime, and include every larger highway by town by date by hour bin cell with at least one accident and at least one stop. This expands our coverage of police stops from 2 to 6 percent. We continue to expand the sample and the cells used for the accident benchmark next to the highway by town by day of week by month by hour bin (19 percent of the sample), then to highway by town by day of week by month by shift (27 percent),[16] highway by town by month by shift (69 percent), and finally, highway by town by shift (93 percent).

As we expand the sample by allowing the accidents to provide a counterfactual over much broader time frames, the disparity estimates using the accident and the telemetric benchmarks remain quite similar, even though the telemetric benchmark continues to be based on the precise time and date of the stop. Notably, the analysis also validates concerns that previous accident benchmark studies might not be representative of overall stop disparities. As we expand the sample, the stop disparities for Black motorists increase substantially. The bottom results contained in Table 8 panel a capture 93 percent of the sample. In these results, the Black disparity using the telemetric benchmark is 4.5 percent and the disparity using the accident benchmark is 5.1 percent, compared to estimates of 2.9 and 3.5 percent for the initial 2 percent sample.

[16] As a reminder, police shifts in Massachusetts are 7am-2pm, 3pm-10pm and 11pm-6am

5.3 Veil of Darkness

Finally, we turn to 'Veil of Darkness' (VOD) estimates, which exploit seasonal variation in the timing of sunset to compare stops made at the same time and location but under differing light conditions. The underlying identification assumption is that during seasons with more daylight, police officers can more easily discern race or ethnicity before making a stop, thereby having the opportunity to discriminate. Conversely, because race is less observable at night, the racial composition of stops made in darkness are assumed to unaffected by officer discrimination and so approximate the composition of motorists on the road who are at risk of being stopped.[17] Standard VOD analyses impose two primary sample restrictions: first, limiting the sample to the 'inter-twilight window' to exclude times that are dark or light year-round; and second, dropping stops for non-moving violations, as the visibility of equipment failures like broken taillights increases substantially in darkness (Grogger and Ridgeway, 2006).

A unique, positive feature of our data—the availability and inclusion of verbal warnings—introduces a specific limitation for this design. Unlike written infractions, verbal warnings in our dataset do not record the reason for the stop, preventing the direct separation of moving from non-moving violations. This problem is complicated by the possibility of disparities in whether police choose to issue verbal warnings, as opposed to written warnings or citations. To address these concerns, we employ two weighting strategies to maintain a representative sample of moving violations. First, we retain verbal warnings but down-weight them based on the observed fraction of written warnings that are moving violations within each highway-by-town-by-shift cell. Alternatively, we drop verbal warnings entirely but up-weight the remaining written warnings to account for the 'lost' volume of verbal stops, using the ratio of total warnings to written warnings within each aggregate cell.

To assess the effect of these sample restrictions, Table 4 re-estimates our racial disparities based on the telemetric benchmark. Panel A replicates our results for the main sample, Panel B repeats the analysis for all stops within the inter-twilight window, Panel C displays the inter-twilight window estimates after removing non-moving written violations while down-weighting verbal warnings, and Panel D presents estimates using written moving violations while up-weighting written warnings. The racial disparities for stops within the inter-twilight window are somewhat larger compared to the full sample of stops in Panel B. However, our reweighted estimates in Panel C and D show that estimated disparities are smaller when non-moving violations are removed from the sample.

Figure 9 presents the VOD estimates and the telemetric benchmark estimates for the inter-twilight sample excluding non-moving violations. The blue lines show the VOD estimates, while the red lines show the telemetric benchmark estimates, and solid and dashed lines identify the samples that retained or dropped the verbal warnings, respectively. Across both moving violation samples and all three racial comparisons, we consistently observe that the VOD estimates of disparity are smaller in magnitude compared to our telemetric benchmark, although this difference is smaller when verbal warnings are dropped. This divergence highlights the conceptual distinction between the two measures. The VOD identification strategy isolates *disparate treatment* by filtering out differences in driving behavior and deployment patterns assuming that those patterns are consistent between daylight and darkness, thereby capturing only discrimination that occurs when the officer can visually discern the motorist's race. In contrast, our telemetric benchmark captures the broader

[17] However, Horrace and Rohlin (2016) argue that race is sometimes visible at night due to artificial lighting, empirically finding that street lights bias the VOD test away from detecting racial disparities. Additionally, Kalinowski et al. (2023) argue that minorities who anticipate discrimination may change their behavior when their race is observable; they show that Black motorists increase the severity of speeding violations and are more likely to be involved in fatal accidents when driving in darkness, ostensibly because minority motorists know that race is harder to observe.

common legal standard of *localized disparate impact*—measuring the total over-representation of minority motorists relative to their actual presence on the roadway. Because the VOD effectively controls for race-neutral structural factors (such as a focus on specific types of violation or the severity of the violation) that contribute to racial disparities regardless of visibility, it systematically yields a lower bound estimate relative to the localized disparate impact experienced by minority motorists, even when conditioning on the locations and times that officers choose to make stops. In fact, our VOD estimates are quite similar to the disparate treatment estimates in police stops against minority Lyft drivers in Aggarwal et al. (2025).

Further, investigation into the mechanisms driving these results shows that when light conditions change, both the share of stops for speeding and the share of stops where warnings were issued increase during daylight hours for all demographic groups. Since the directional change in enforcement behavior is consistent across all groups, the observed shifts are unlikely to be driven solely by changes in the racial composition of motorists. In fact, we rule out motorist composition changes by showing that the telemetric composition of motorists on the road is not correlated with daylight.

6 Conclusion

This paper addresses the fundamental challenge of establishing a valid counterfactual for evaluating localized disparate impact in police traffic stops. By leveraging high-frequency telemetric mobility data calibrated with administrative crash records, we construct a dynamic measure of the driving population that accounts for the complex spatiotemporal variation in determining who is at risk of being stopped. Applying this measure to traffic stops by the Massachusetts State Police, we find robust evidence of localized disparate impact against minority motorists. Our granular estimates reveal that Black and Hispanic motorists are stopped at rates exceeding their presence on the roadway by 4.4 and 4.9 percentage points, respectively. Further, our estimates demonstrate the importance of using an outside benchmark for calibrating telemetric data on race. Specifically, using just telemetric data, minority motorists comprise only 15.3 percent of the estimated driving population around traffic stops relative to comprising 34.8 percent of the stops, which implies a large 19.5 percent disparity comparable to stop disparity estimates using vendor provided telemetric based racial composition (Xu et al., 2024), but after calibrating the telemetric predicted share to the accident share minority the estimate of minority motorist share is 27.71 percent of the driving population implying minority disparities of only 7.1 percent. As discussed above, the larger estimates in Xu et al. (2024) may arise in part due to the inability of their data to capture higher minority shares of motorists outside of regular commuting times or may capture a broader measure of disparate impact arising from decisions about when and where police focus on traffic enforcement, such as the locational disparities documented in speeding enforcement by Cai et al. (2022).

Beyond documenting stop disparities relative to the common legal standard of the minority share of motorists, our study provides a rigorous evaluation of the metrics currently employed in or considered for policy and litigation. We demonstrate that the widely used Community Standard benchmark, which relies on residential Census data, is structurally flawed for assessing traffic enforcement, even after sample adjustments that should reduce bias from commuting and longer distance travel patterns. Disparity estimates are inflated by 7 to 13 percentage points relative to our estimate of the composition of the roadway, and dropping stops during rush hour or on major interstates has minimal effect on this bias. Conversely, we find that the Veil of Darkness (VOD) test yields estimates that are consistently smaller than our telemetric benchmark. This divergence confirms that while VOD is a powerful tool for isolating disparate *treatment* (disparities dependent on visibility), it understates the broader legal standard of localized disparate *impact*, effectively

screening out structural factors that result in the over-policing of minority motorists regardless of lighting conditions. Notably, we find minimal evidence of changes in the racial composition of motorists in darkness, but substantial changes in policing behavior even when excluding non-moving violations.

Perhaps the most significant finding for policymakers is the validation of the Crash Benchmark. Historically viewed with skepticism due to the relative rarity of motor vehicle accidents, the crash benchmark performs remarkably well when validated against our telemetric measure. We show that when the crash benchmark is extrapolated to higher levels of aggregation — specifically the highway-by-town-by-shift level — estimates remain similar to the telemetric benchmark that uses the precise date and time of the stop, while expanding coverage from 2 percent to 93 percent of all police stops. Notably, the correspondence between the crash and the telemetric benchmark estimates is preserved even though the point estimates of disparities increases substantially as the coverage of the crash benchmark is expanded, i.e. Black motorist disparities rise from 3.5 to 5.1 percentage points using the crash benchmark. This finding offers a practical solution for jurisdictions that lack the resources to acquire and process commercial telemetric data.

Effective governance requires accurate measurement. As state statutes and federal consent decrees increasingly mandate the monitoring of minority stop share relative to share of motorists, what we refer to as localized disparate impact, as the estimand of interest, the choice of benchmark dictates the conclusions drawn about police conduct. Our results indicate that while telemetric data offers a potential "gold standard" for measurement, reliable use of telemetric data requires calibration to some external benchmark for the minority share of motorists, like accident or perhaps traffic camera data, because race is not observed in telemetric data. Further, administrative accident data by itself provides a cost-effective and accurate alternative that can be expanded to cover the vast majority of traffic stops and far surpasses the reliability of residential population comparisons. By adopting these empirically validated benchmarks, policymakers can move beyond debates over measurement methodology and focus on the substantive work of addressing the drivers of racial disparities in the criminal justice system.

References

Abrahams, S. (2020, December). Officer differences in traffic stops of minority drivers. *Labour Economics 67*, 101912.

Aggarwal, P., A. Brandon, A. Goldszmidt, J. Holz, J. A. List, I. Muir, G. Sun, and T. Yu (2025). High-frequency location data show that race affects citations and fines for speeding. *Science 387*(6741), 1397–1401.

Alpert, G. P., M. R. Smith, and R. G. Dunham (2004, June). Toward a better benchmark: Assessing the utility of not-at-fault traffic crash data in racial profiling research. *Justice Res. Policy 6*(1), 43–69.

Ang, D., P. Bencsik, J. Bruhn, and E. Derenoncourt (2021, September). Police violence reduces civilian cooperation and engagement with law enforcement. *HKS Working Paper* (RWP21-022).

Anwar, S. and H. Fang (2006). An alternative test of racial prejudice in motor vehicle searches: Theory and evidence. *American Economic Review 96*(1), 127–151.

Athey, S., B. Ferguson, M. Gentzkow, and T. Schmidt (2021). Estimating experienced racial segregation in us cities using large-scale gps data. *Proceedings of the National Academy of Sciences 118*.

Bunten, D. M., E. Fu, L. Rolheiser, and C. Severen (2024, May). The problem has existed over endless years: Racialized difference in commuting, 1980–2019. *Journal of Urban Economics 141*, 103542.

Bushway, S., A. Jordan, D. Neal, and S. Raphael (2025). Understanding racial disparities in criminal court outcomes. *RSF Journal 11*, 86–135.

Cai, W., J. Gaebler, J. Kaashoek, L. Pinals, S. Madden, and S. Goel (2022, 07). Measuring racial and ethnic disparities in traffic enforcement with large-scale telematics data. *PNAS Nexus 1*(4), pgac144.

Chen, M. K., K. L. Christensen, E. John, E. Owens, and Y. Zhuo (2025, 11). Smartphone data reveal neighborhood-level racial disparities in police presence. *The Review of Economics and Statistics 107*(6), 1734–1742.

Couture, V., J. I. Dingel, A. E. Green, and J. Handbury (2025). Demographic preferences and income segregation. *Working Paper*.

Dunn, R. A. (2014). Measuring racial disparities in traffic ticketing within large urban jurisdictions. *Public Performance and Management Review*, 537–561. Published online 08 December 2014.

Feigenberg, B. and C. Miller (2022). Would eliminating racial disparities in motor vehicle searches have efficiency costs? *Quarterly Journal of Economics 137*, 49–113.

Goncalves, F. and S. Mello (2021). A few bad apples? racial bias in policing. *American Economic Review 111*, 1406–1441.

Gong, X., Y. Han, S. T. Parker, M. B. Ross, and S. L. Ross (2025). Are police officers monolithic in their treatment of minority motorists - how heterogeneous are officer traffic stop patterns over race? Technical report.

Green, A., M. Kutzbach, and L. Vilhuber (2017). Two perspectives on commuting: A comparison of home to work flows across job-linked survey and administrative files. *Working Paper*.

Grogger, J. and G. Ridgeway (2006). Testing for racial profiling in traffic stops from behind a veil of darkness. *Journal of the American Statistical Association 101*(475), 878–887.

Horrace, W. C. and S. M. Rohlin (2016, May). How dark is dark? bright lights, big city, racial profiling. *The Review of Economics and Statistics 98*(2), 226–232.

Jardel, H. V. and P. L. Delamater (2024, April). Uncovering representation bias in large-scale cellular phone-based data: A case study in north carolina. *Geographical Analysis*. First published 08 April 2024.

Kalinowski, J. J., M. B. Ross, and S. L. Ross (2023). Endogenous driving behavior in tests of racial profiling. *Journal of Human Resources*.

Kalinowski, J. J., M. B. Ross, and S. L. Ross (2025). Endogenous driving behavior in tests of racial profiling. *Journal of Human Resources*.

Knowles, J., N. Persico, and P. Todd (2001, February). Racial bias in motor vehicle searches: Theory and evidence. *Journal of Political Economy 109*(1), 203–229.

Li, Z., H. Ning, F. Jing, and M. N. Lessani (2024, January). Understanding the bias of mobile location data across spatial scales and over time: A comprehensive analysis of SafeGraph data in the united states. *PLoS One 19*(1), e0294430.

Massachusetts Department of Transportation (2012, June). Massachusetts travel survey. Technical report, Massachusetts Department of Transportation. Accessed 2025.

Mello, S. (2025, October). Fines and financial wellbeing. *The Review of Economic Studies 92*(5), 3340–3374.

Pierson, E., C. Simoiu, J. Overgoor, S. Corbett-Davies, D. Jenson, A. Shoemaker, V. Ramachandran, P. Barghouty, C. Phillips, R. Shroff, and S. Goel (2020, July). A large-scale analysis of racial disparities in police stops across the united states. *Nature Human Behaviour 4*(7), 736–745.

Ridgeway, G. and J. M. MacDonald (2009). 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.

Ridgeway, G. and J. M. MacDonald (2014). A method for internal benchmarking of criminal justice system performance. *Crime and Delinquency 60*(1), 145–162.

Rojek, J., R. Rosenfeld, and S. Decker (2004). The influence of driver's race on traffic stops in missouri. *Police Quarterly 7*(1), 63–90.

Smith, M. R. and M. Petrocelli (2001). Racial profiling? a multivariate analysis of police traffic stop data. *Police Quarterly 4*(1), 4–27.

Tapp, S. N. and E. Davis (2024, October). Contacts between police and the public, 2022. Technical Report NCJ 308847, Bureau of Justice Statistics, U.S. Department of Justice.

Walker, S. (2022). The justice department's pattern-or-practice police reform program, 1994–2017: Goals, achievements, and issues. *Annual Review of Criminology 5*(1), 21–42.

Warren, P., D. Tomaskovic-Devey, W. Smith, M. Zingraff, and M. Mason (2006). Driving while black: Bias processes and racial disparity in police stops. *Criminology 44*(3), 709–738.

Weisburd, S. (2021, 05). Police presence, rapid response rates, and crime prevention. *The Review of Economics and Statistics 103*(2), 280–293.

Weisburst, E. K. (2022). "whose help is on the way?" the importance of individual police officers in law enforcement outcomes. *The Journal of human resources*. 0720-11019R2.

West, J. (2018). Racial bias in police investigations. *Working paper*.

Xu, W., M. Smart, N. Tilahun, S. Askari, Z. Dennis, H. Li, and D. Levinson (2024). The racial composition of road users, traffic citations, and police stops. *Proceedings of the National Academy of Sciences 121*(24), e2402547121.

Table 1: First Stage Regression

	Black		Hispanic		White	
	(1) Sample 1	(2) Sample 2	(3) Sample 1	(4) Sample 2	(5) Sample 1	(6) Sample 2
Sample 2	0.284***		0.230***		0.287***	
	(0.0215)		(0.0168)		(0.0164)	
Sample 1		0.323***		0.238***		0.300***
		(0.0216)		(0.0181)		(0.0173)
N	63730	63730	65692	65692	76619	76619
F	174.7528	224.2085	187.2118	173.975	307.6252	299.9934
p	3.35e-37	2.24e-46	1.46e-39	4.54e-37	2.65e-61	5.52e-60

Standard errors in parentheses, cluster at town × route × shift level
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2: Second Stage Regression

	Black		Hispanic		White	
	(1) Accident	(2) Accident	(3) Accident	(4) Accident	(5) Accident	(6) Accident
Sample 1	2.206***		0.829**		1.350***	
	(0.338)		(0.308)		(0.201)	
Sample 2		2.543***		1.339***		1.607***
		(0.334)		(0.300)		(0.207)
FE Mean	0.0147	-0.0090	0.0971	0.0355	-0.4245	-0.6356
N	63730	63730	65692	65692	76619	76619
F	42.60702	57.80268	7.224442	19.95341	45.13317	60.19441
p	1.01e-10	6.07e-14	.0072968	8.73e-06	2.87e-11	1.87e-14

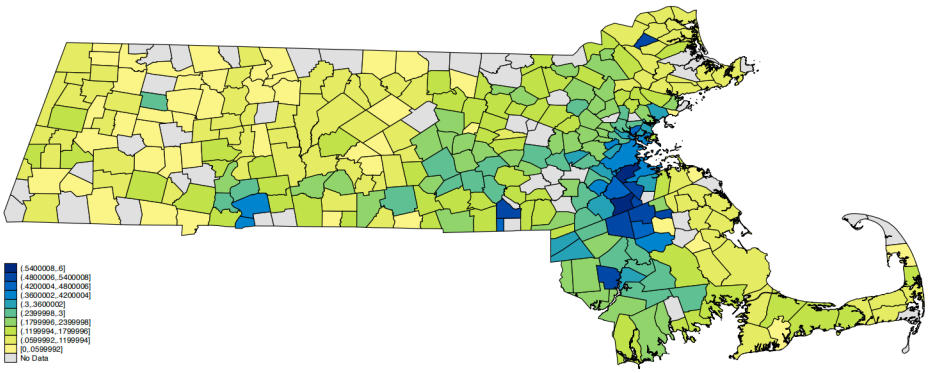
Standard errors in parentheses, cluster at town × route × shift level
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3: Weighted Standard Deviation Summary

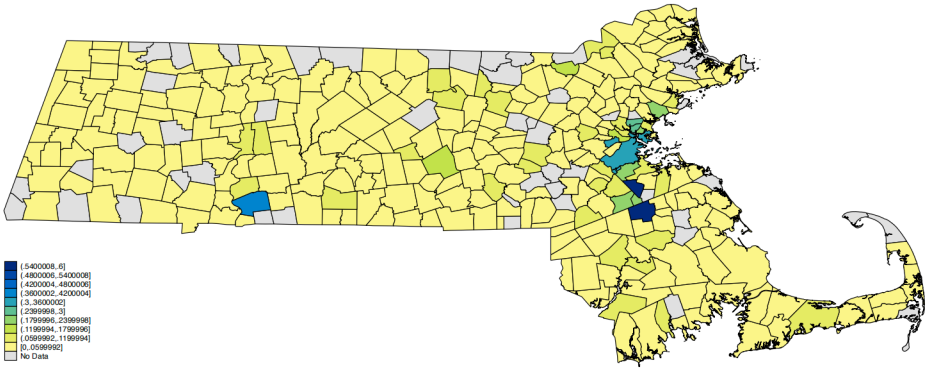
	(1) Black	(2) Hispanic	(3) White
Panel A: Town Level			
Stops - Mean	0.104447	0.110962	0.133937
Stops - Telemetric Benchmark	0.055345	0.053901	0.057477
Stops - Ratio in Census	0.083662	0.102309	0.130771
Panel B: Hour of Day by Day of Week			
Stops - Mean	0.060382	0.043550	0.066438
Stops - Telemetric Benchmark	0.026518	0.020010	0.023333
Panel B: Week of Year Level			
2021 Stops - Mean	0.013312	0.014062	0.016962
2021 Stops - Telemetric Benchmark	0.010940	0.009490	0.010876
2022 Stops - Mean	0.013151	0.013575	0.016212
2022 Stops - Telemetric Benchmark	0.009926	0.011078	0.011976

Figure 1: Geographic Variation in Black Traffic Stops, Census Data, and the Telemetric Benchmark

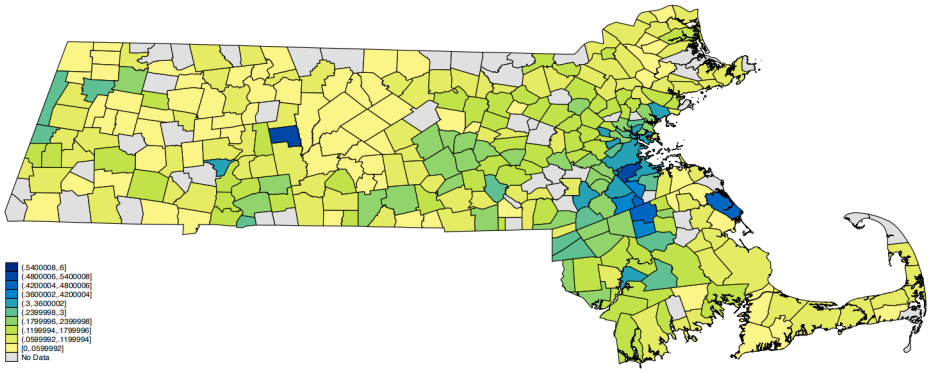
(a) Black Traffic Stops



(b) Black Ratio in Census



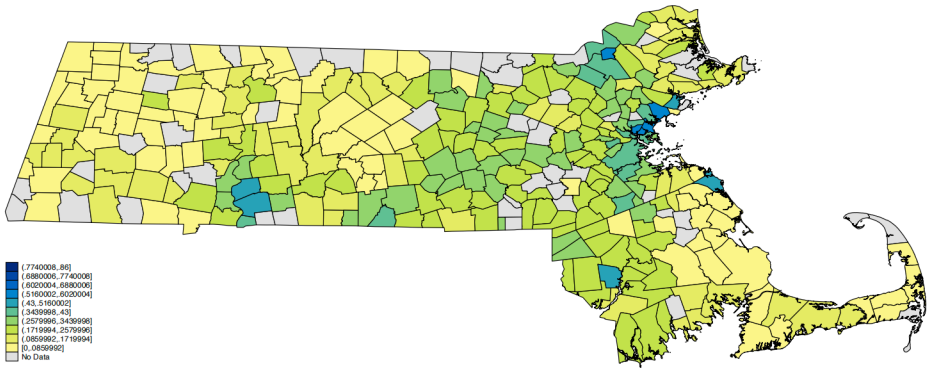
(c) Black Telemetric Benchmark



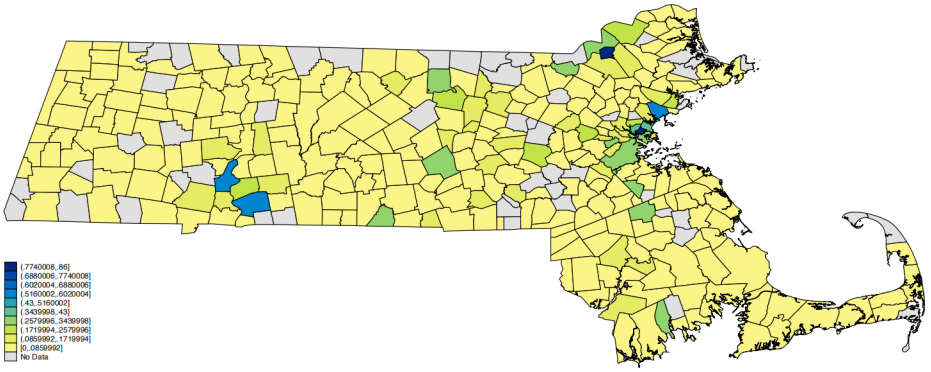
*N*otes: The Census data used are from the 2020 American Community Survey 5-Year Estimates. Use day of week, day of month, month and route fixed effects.

Figure 2: Geographic Variation in Hispanic Traffic Stops, Census Data, and the Telemetric Benchmark

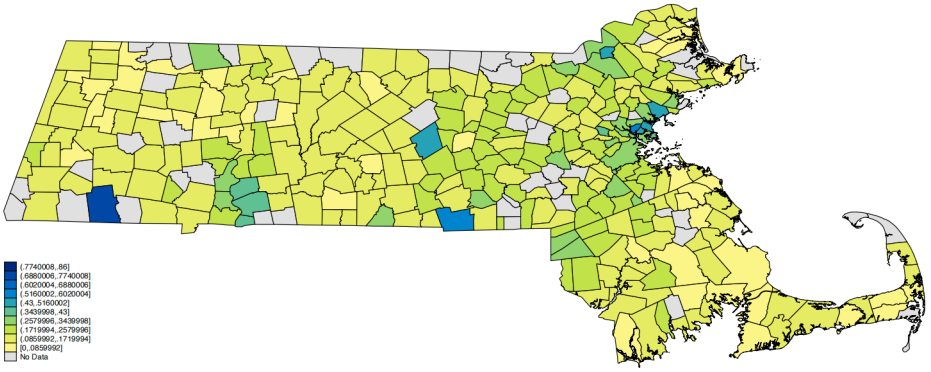
(a) Hispanic Traffic Stops



(b) Hispanic Ratio in Census



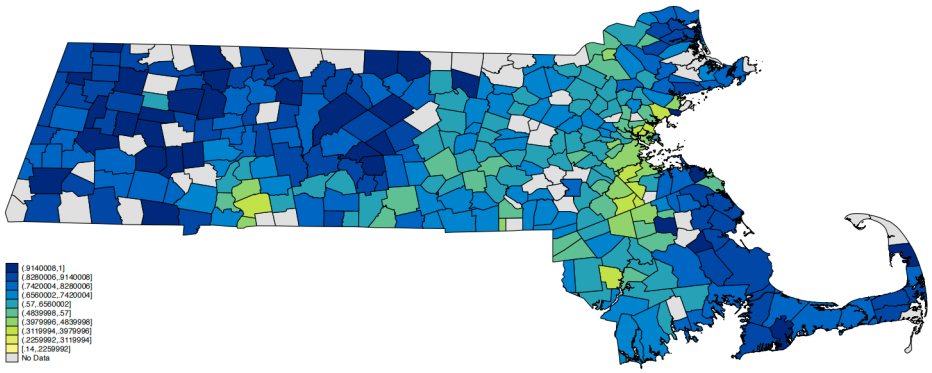
(c) Hispanic Telemetric Benchmark



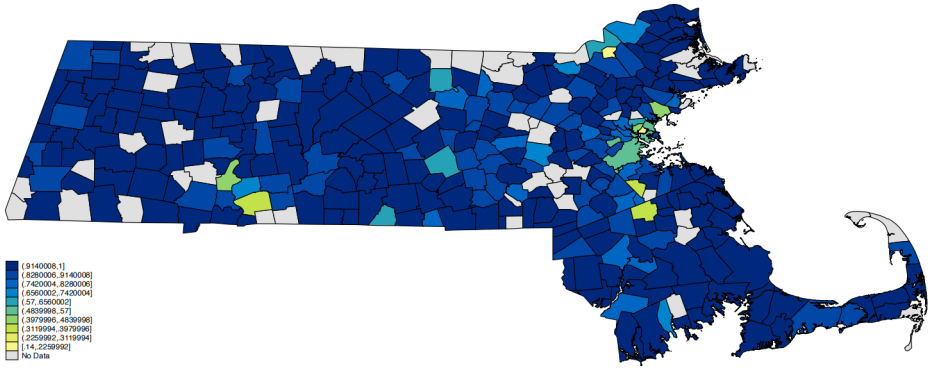
*N*otes: The Census data used are from the 2020 American Community Survey 5-Year Estimates. Use day of week, day of month, month and route fixed effects.

Figure 3: Geographic Variation in White Traffic Stops, Census Data, and the Telemetric Benchmark

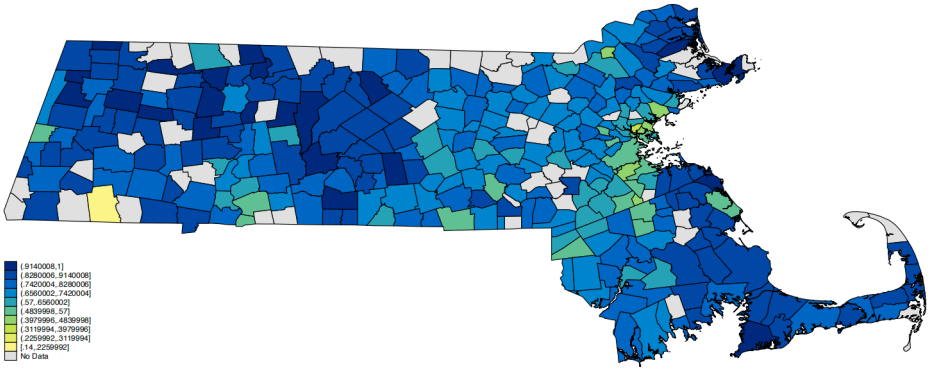
(a) White Traffic Stops



(b) White Ratio in Census



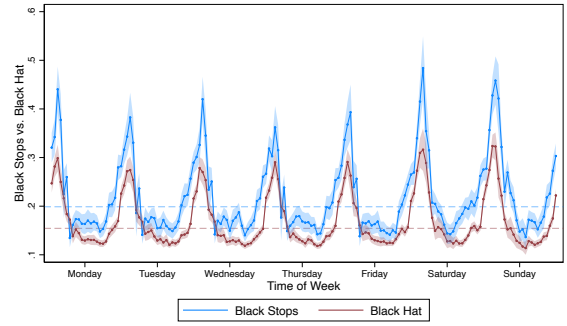
(c) White Telemetric Benchmark



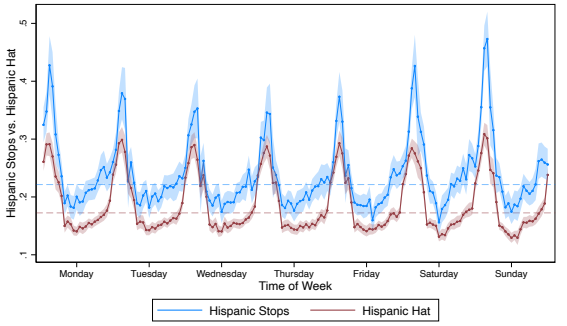
*N*otes: The Census data used are from the 2020 American Community Survey 5-Year Estimates. Use day of week, day of month, month and route fixed effects.

Figure 4: Temporal Variation in Traffic Stops and the Telemetric Benchmark

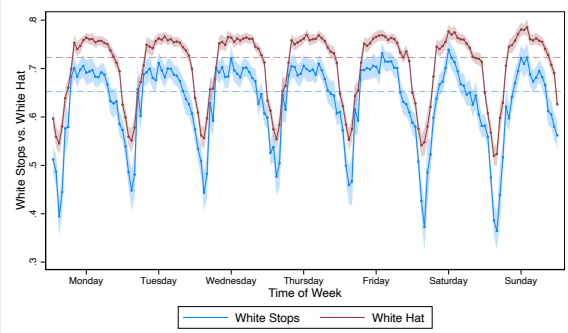
(a) Black Traffic Stops vs. Telemetric Benchmark



(b) Hispanic Traffic Stops vs. Telemetric Benchmark



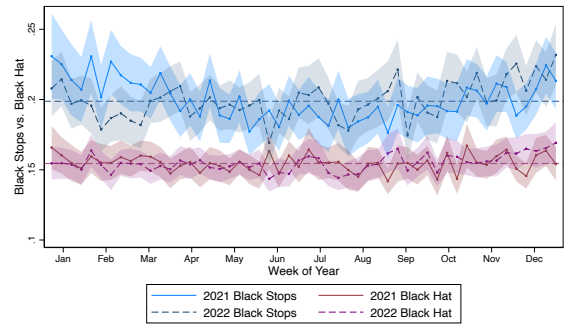
(c) White Traffic Stops vs. Telemetric Benchmark



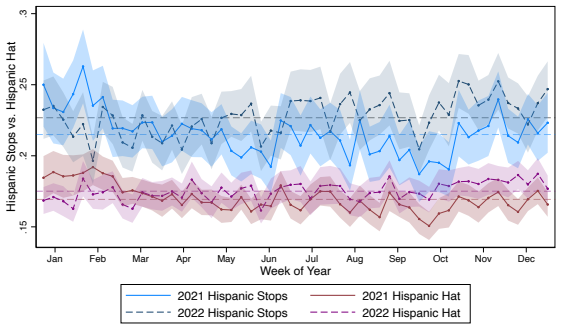
*N*otes: Use day of month, month and town × route fixed effects.

Figure 5: Seasonal Variation in Traffic Stops and the Telemetric Benchmark

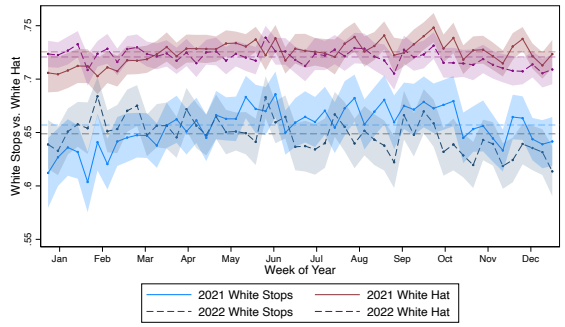
(a) Black Traffic Stops vs. Telemetric Benchmark



(b) Hispanic Traffic Stops vs. Telemetric Benchmark



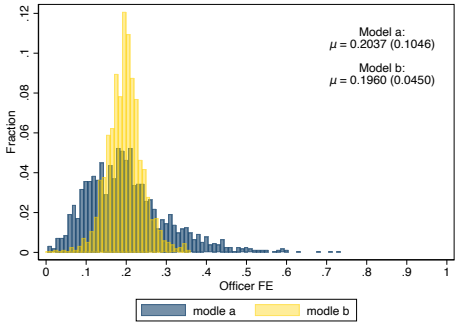
(c) White Traffic Stops vs. Telemetric Benchmark



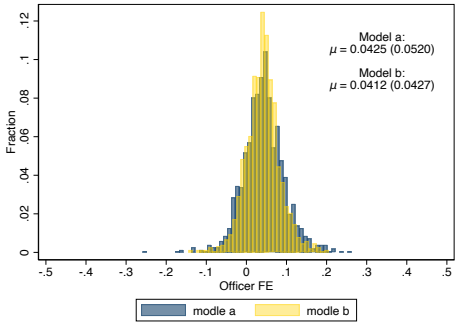
*N*otes: Use day of week, day of month and town × route fixed effects.

Figure 6: Distribution of Estimated Officer Stop Propensities

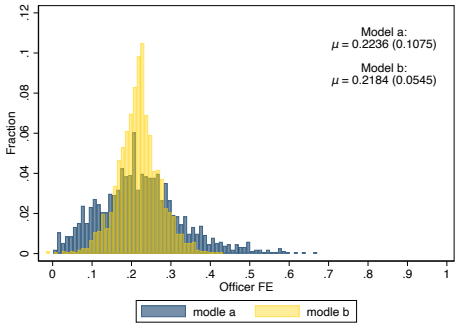
(a) Black Traffic Stops



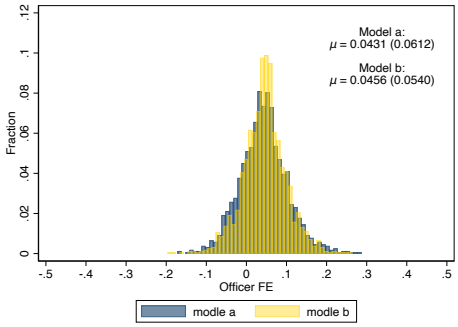
(b) Black Stops - Black Telemetric Benchmark



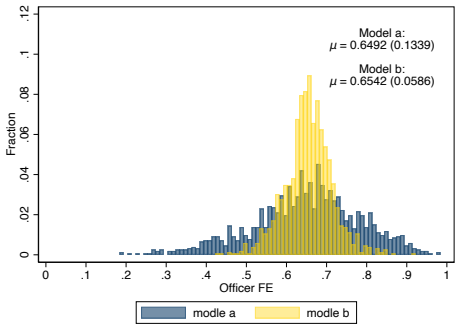
(c) Hispanic Traffic Stops



(d) Hispanic Stops - Hispanic Telemetric Benchmark



(e) White Traffic Stops



(f) White Stops - White Telemetric Benchmark

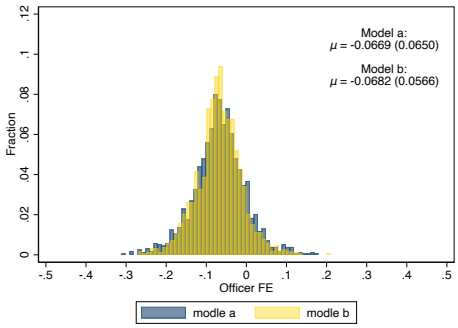
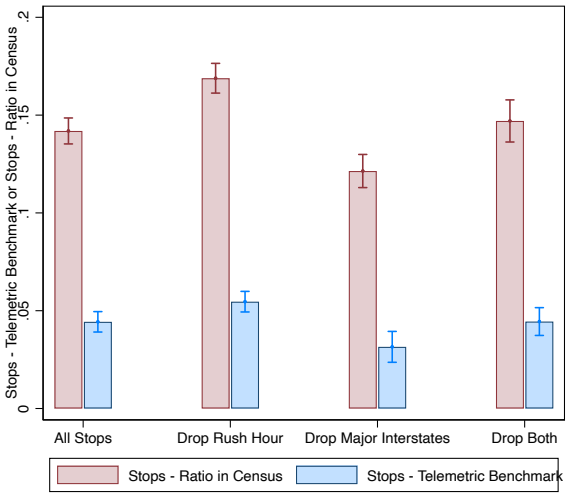
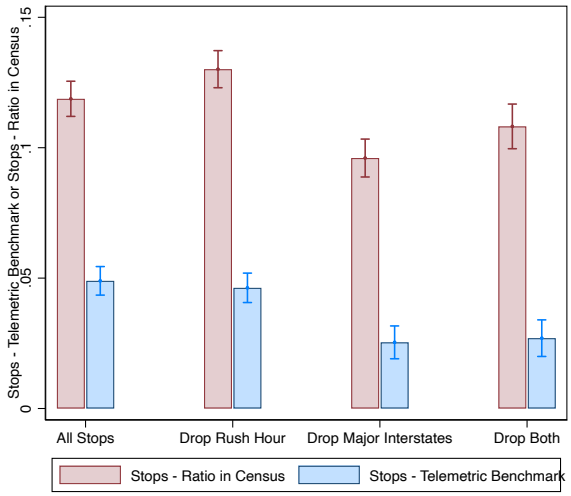


Figure 7: Community Standard

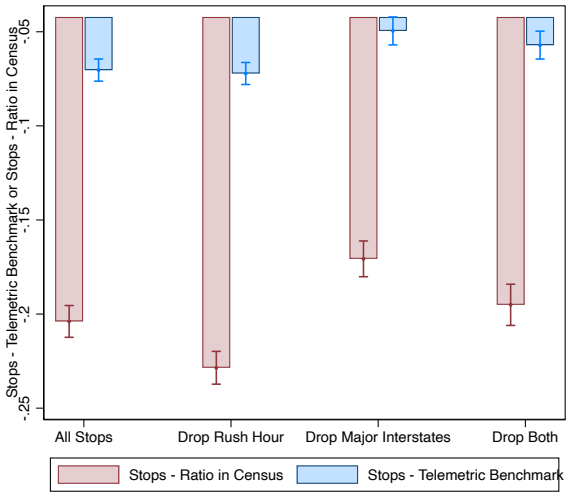
(a) Black



(b) Hispanic



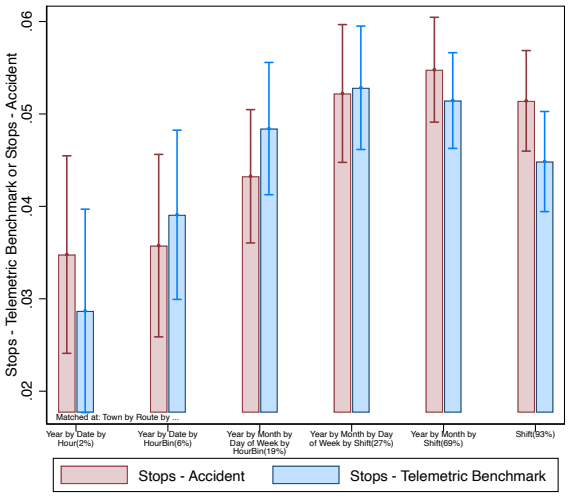
(c) White



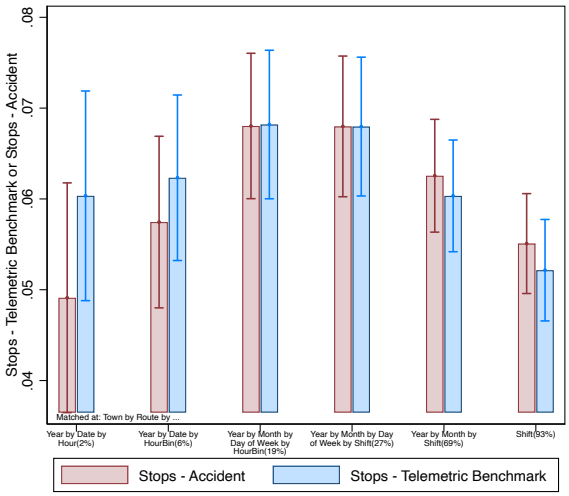
*N*otes: Standard errors in parentheses, cluster at officer level
Rush Hour: 6am-10am and 4pm-8pm
Major Interstates: I-84, I-90, I-91, I-95, I-93, I-190, I-195, I-290, I-395, I-495

Figure 8: Crash / Accident Data

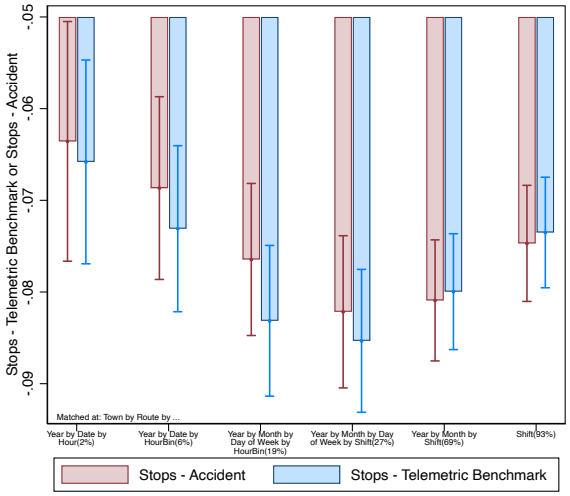
(a) Black



(b) Hispanic



(c) White



*N*otes: Standard errors in parentheses, cluster at officer level
T is town, R is route, Y is year, M is month and DoW is day of week
Hour bin: 2am-5am, 6am-10am, 11am-3pm, 4pm-8pm, 9pm-1am
Shift: 7am-2pm, 3pm-10pm, 11pm-6am

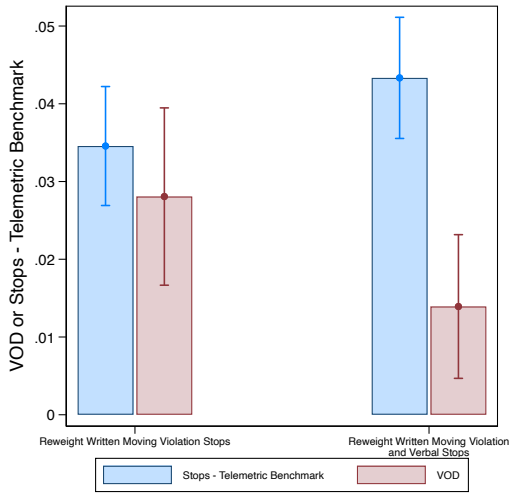
Table 4: Stops - Telemetric Benchmark

	(1) Black	(2) Hispanic	(3) White
Panel A: All Stops			
Weighted Average	0.0443	0.0490	-0.0704
	(0.00266)	(0.00279)	(0.00300)
N	268361	275326	322783
Panel B: Stops Fall Within Interwilight Window			
Weighted Average	0.0484	0.0510	-0.0769
	(0.00306)	(0.00371)	(0.00399)
N	67618	71156	82137
Panel C: Reweight Written Moving Violation and Verbal Stops Fall Within Interwilight Window			
Weighted Average	0.0432	0.0401	-0.0677
	(0.00331)	(0.00344)	(0.00382)
N	51540	53706	62327
Panel D: Reweight Written Moving Violation Stops Fall Within Interwilight Window			
Weighted Average	0.0341	0.0336	-0.0575
	(0.00319)	(0.00371)	(0.00410)
N	32302	33763	38435

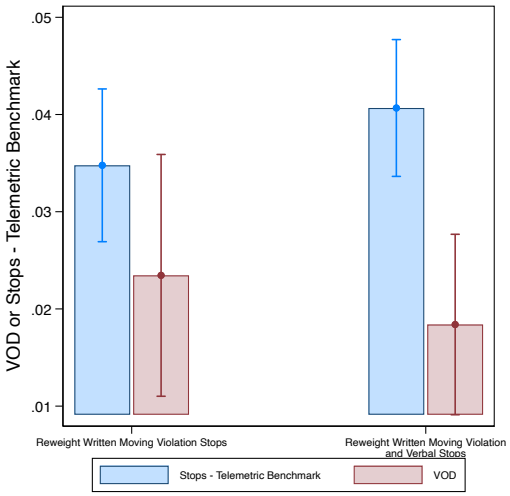
Standard errors in parentheses, cluster at officer level

Figure 9: Veil of Darkness

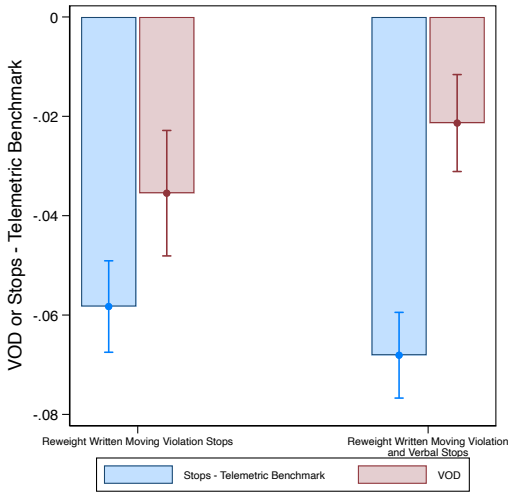
(a) Black



(b) Hispanic



(c) White



Notes: Standard errors cluster at town by route by hour level

Table 5: Veil of Darkness

	Veil of Darkness			Stops - Telemetric Benchmark		
	(1) Black	(2) Hispanic	(3) White	(4) Black	(5) Hispanic	(6) White
Panel A: Reweight Written Moving Violation and Verbal Stops						
disparity	0.0139**	0.0184***	-0.0214***	0.0433***	0.0407***	-0.0681***
	(0.00471)	(0.00474)	(0.00498)	(0.00398)	(0.00359)	(0.00440)
Mean of Stops	0.1810	0.2145	0.6693	0.1810	0.2145	0.6693
town × route × hour	Y	Y	Y			
officer	Y	Y	Y			
N	50738	52951	61596	50738	52951	61596
Panel B: Reweight Written Moving Violation Stops						
disparity	0.0281***	0.0235***	-0.0355***	0.0346***	0.0348***	-0.0583***
	(0.00582)	(0.00635)	(0.00644)	(0.00390)	(0.00401)	(0.00470)
Mean of Stops	0.1643	0.2055	0.6871	0.1643	0.2055	0.6871
town × route × hour	Y	Y	Y			
officer	Y	Y	Y			
N	31413	32892	37594	31413	32892	37594

Standard errors in parentheses, cluster at town × route × hour level

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Hat utilizes an IV constructed from accidents occurring within the inter-twilight window

APPENDICES

A Supplementary results

Table A-1: Correlation Matrix

Panel A: Black			
	Stops - Mean	Stops - Ratio in Census	Stops - Telemetric Benchmark
Stops - Mean	1	.6701689	.8490066
Stops - Ratio in Census	.6701689	1	.6205666
Stops - Telemetric Benchmark	.8490066	.6205666	1
Panel B: Hispanic			
	Stops - Mean	Stops - Ratio in Census	Stops - Telemetric Benchmark
Stops - Mean	1	.6616647	.8748514
Stops - Ratio in Census	.6616647	1	.7119293
Stops - Telemetric Benchmark	.8748514	.7119293	1
Panel C: White			
	Stops - Mean	Stops - Ratio in Census	Stops - Telemetric Benchmark
Stops - Mean	1	.6380948	.9033258
Stops - Ratio in Census	.6380948	1	.6552804
Stops - Telemetric Benchmark	.9033258	.6552804	1

Table A-2: Correlation Matrix

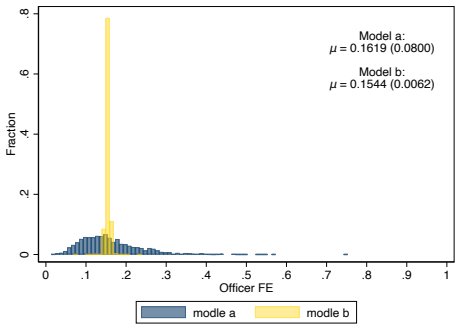
Panel A: Black		
	Stops	Telemetric Benchmark
Stops	1	.9269835
Telemetric Benchmark	.9269835	1
Panel B: Hispanic		
	Stops	Telemetric Benchmark
Stops	1	.888359
Telemetric Benchmark	.888359	1
Panel C: White		
	Stops	Telemetric Benchmark
Stops	1	.9377741
Telemetric Benchmark	.9377741	1

Table A-3: Correlation Matrix

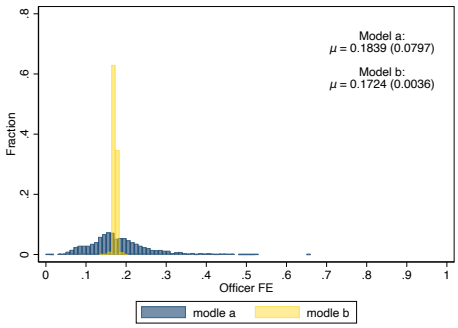
Panel A: Black		
	2021 Stops	2021 Telemetric Benchmark
2021 Stops	1	.5876578
2021 Telemetric Benchmark	.5876578	1
Panel B: Hispanic		
	2021 Stops	2021 Telemetric Benchmark
2021 Stops	1	.7535257
2021 Telemetric Benchmark	.7535257	1
Panel C: White		
	2021 Stops	2021 Telemetric Benchmark
2021 Stops	1	.8051168
2021 Telemetric Benchmark	.8051168	1
Panel A: Black		
	2022 Stops	2022 Telemetric Benchmark
2022 Stops	1	.7120308
2022 Telemetric Benchmark	.7120308	1
Panel B: Hispanic		
	2022 Stops	2022 Telemetric Benchmark
2022 Stops	1	.5921804
2022 Telemetric Benchmark	.5921804	1
Panel C: White		
	2022 Stops	2022 Telemetric Benchmark
2022 Stops	1	.7265656
2022 Telemetric Benchmark	.7265656	1

Figure A-1: Distribution of Estimated Officer Telemetric Benchmark

(a) Black



(b) Hispanic



(c) White

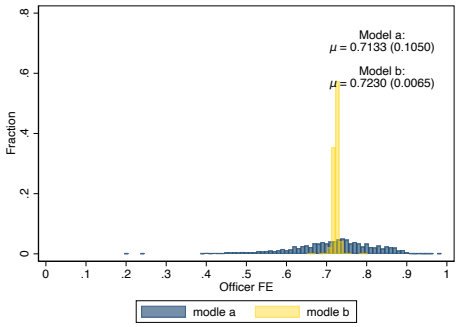


Table A-4: Correlation Matrix of Black

	Stops Model A	Stops Model B	Telemetric Model A	Telemetric Model B	Stops - Telemetric Model A	Stops - Telemetric Model B
Stops Model A	1	.7332187	.7651897	.146719	.7690326	.7236385
Stops Model B	.7332187	1	.2953491	.2013017	.8680986	.9934983
Telemetric Model A	.7651897	.2953491	1	.1853337	.230517	.2770257
Telemetric Model B	.146719	.2013017	.1853337	1	.1044717	.1286831
Stops - Telemetric Model A	.7690326	.8680986	.230517	.1044717	1	.8732183
Stops - Telemetric Model B	.7236385	.9934983	.2770257	.1286831	.8732183	1

Table A-5: Correlation Matrix of Hispanic

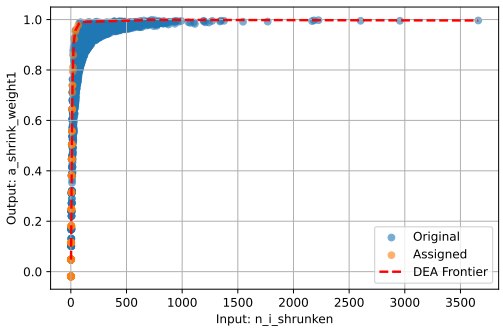
	Stops Model A	Stops Model B	Telemetric Model A	Telemetric Model B	Stops - Telemetric Model A	Stops - Telemetric Model B
Stops Model A	1	.7526469	.7175179	.1430108	.7834064	.7441662
Stops Model B	.7526469	1	.24755	.0821756	.8822284	.9979635
Telemetric Model A	.7175179	.24755	1	.1751559	.1759216	.2361694
Telemetric Model B	.1430108	.0821756	.1751559	1	.0695893	.039977
Stops - Telemetric Model A	.7834064	.8822284	.1759216	.0695893	1	.8813973
Stops - Telemetric Model B	.7441662	.9979635	.2361694	.039977	.8813973	1

Table A-6: Correlation Matrix of White

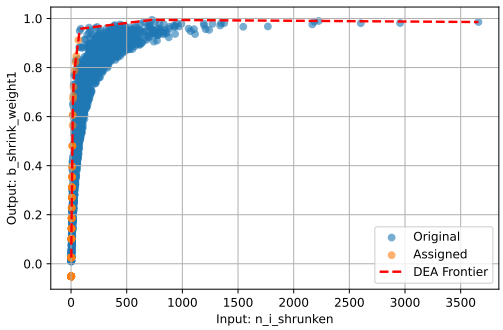
	Stops Model A	Stops Model B	Telemetric Model A	Telemetric Model B	Stops - Telemetric Model A	Stops - Telemetric Model B
Stops Model A	1	.7461718	.7849875	.2183853	.7362344	.7314153
Stops Model B	.7461718	1	.2999542	.1876555	.9035215	.9945939
Telemetric Model A	.7849875	.2999542	1	.2759806	.2069229	.2767309
Telemetric Model B	.2183853	.1876555	.2759806	1	.104041	.1135656
Stops - Telemetric Model A	.7362344	.9035215	.2069229	.104041	1	.9080865
Stops - Telemetric Model B	.7314153	.9945939	.2767309	.1135656	.9080865	1

Figure A-2: Bayesian Shrinkage Weights for Officers and DEA Frontier, Black Stops

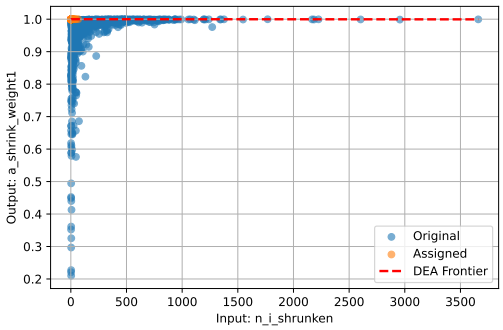
(a) Black Stops, Model A



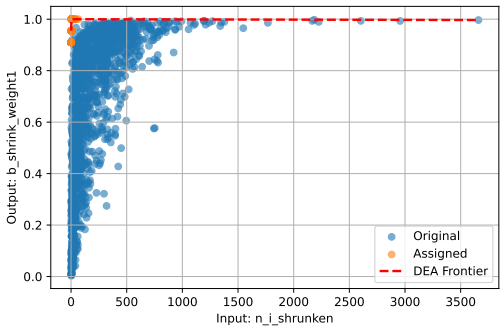
(b) Black Stops, Model B



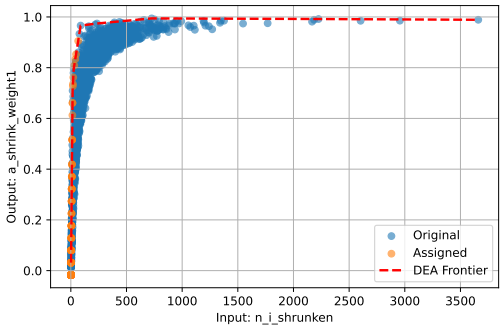
(c) Black Telemetric Benchmark, Model A



(d) Black Telemetric Benchmark, Model B



(e) Black Stops - Telemetric Benchmark, Model A



(f) Black Stops - Telemetric Benchmark, Model B

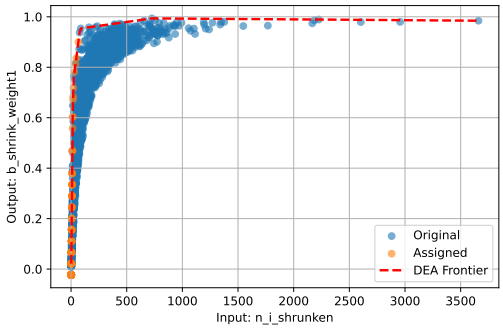
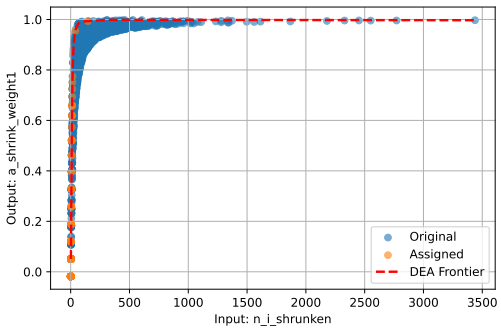
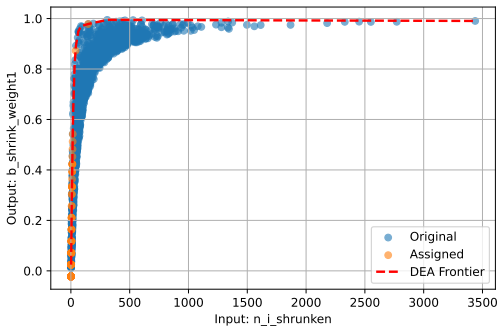


Figure A-3: Bayesian Shrinkage Weights for Officers and DEA Frontier, Hispanic Stops

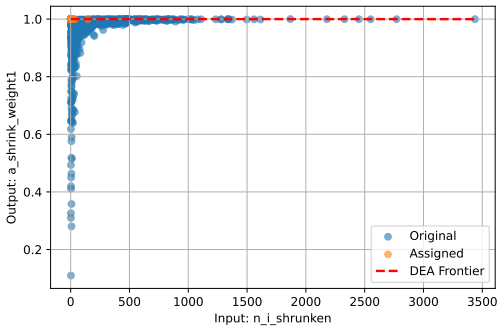
(a) Hispanic Stops, Model A



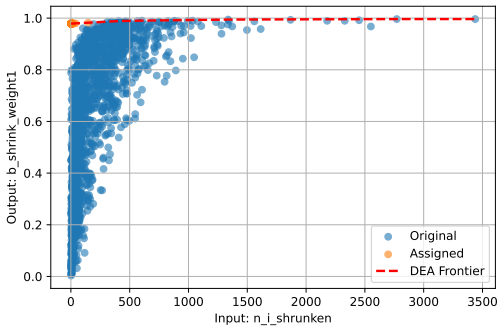
(b) Hispanic Stops, Model B



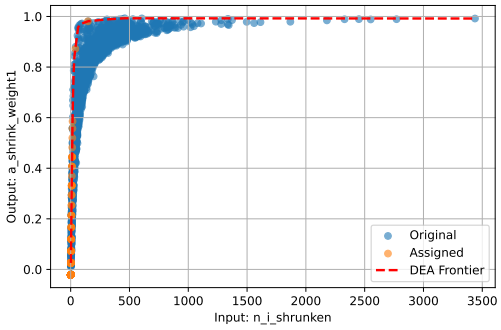
(c) Hispanic Telemetric Benchmark, Model A



(d) Hispanic Telemetric Benchmark, Model B



(e) Hispanic Stops - Telemetric Benchmark, Model A



(f) Hispanic Stops - Telemetric Benchmark, Model B

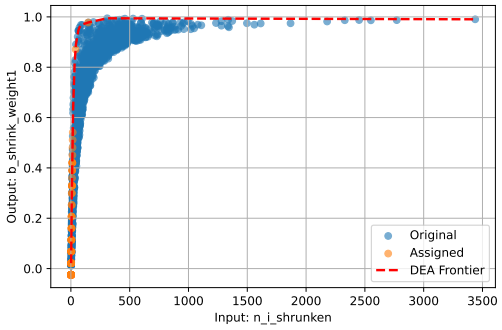
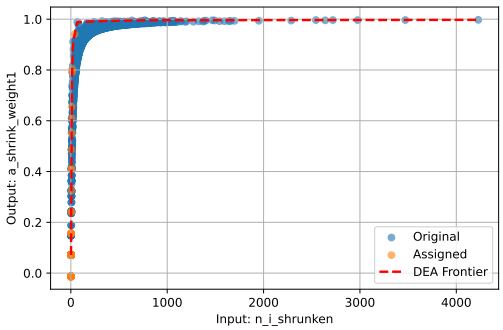
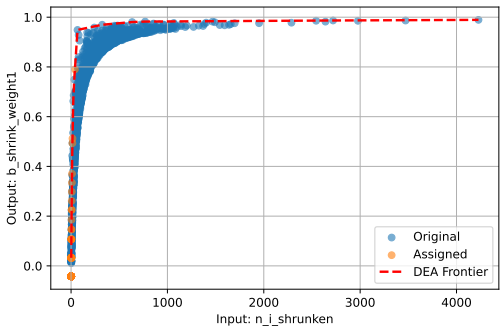


Figure A-4: Bayesian Shrinkage Weights for Officers and DEA Frontier, White Stops

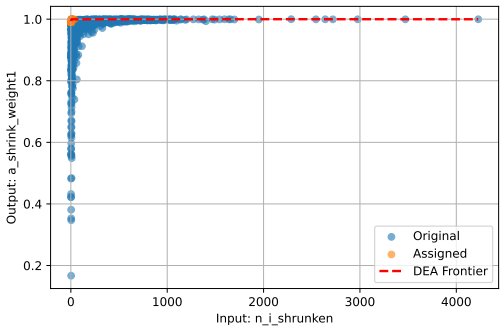
(a) White Stops, Model A



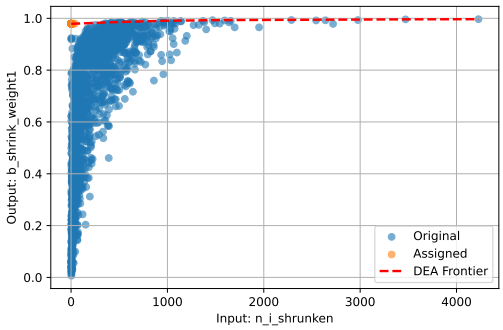
(b) White Stops, Model B



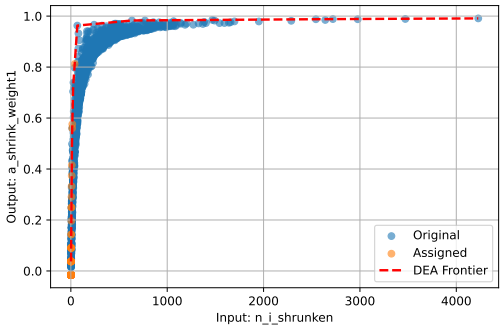
(c) White Telemetric Benchmark, Model A



(d) White Telemetric Benchmark, Model B



(e) White Stops - Telemetric Benchmark, Model A



(f) White Stops - Telemetric Benchmark, Model B

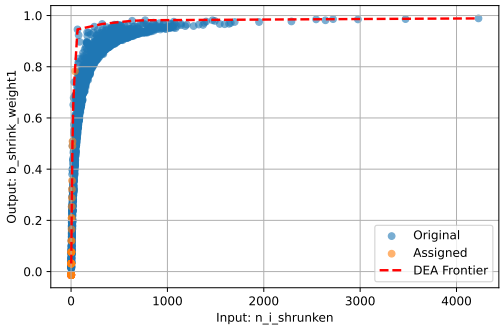


Table A-7: Community Standard

	Stops - Telemetric Benchmark			Stops - Ratio in Census		
	(1) Black	(2) Hispanic	(3) White	(4) Black	(5) Hispanic	(6) White
All Stops						
Weighted Average	0.0443	0.0490	-0.0704	0.142	0.119	-0.204
	(0.00266)	(0.00279)	(0.00300)	(0.00338)	(0.00344)	(0.00431)
N	268361	275326	322783	268361	275326	322783
Drop Rush Hour						
Weighted Average	0.0546	0.0463	-0.0722	0.169	0.130	-0.229
	(0.00269)	(0.00288)	(0.00299)	(0.00388)	(0.00363)	(0.00444)
N	139160	140404	169202	139160	140404	169202
Drop Major Interstates						
Weighted Average	0.0315	0.0254	-0.0496	0.121	0.0960	-0.171
	(0.00403)	(0.00321)	(0.00379)	(0.00432)	(0.00371)	(0.00485)
N	113429	114877	133924	113429	114877	133924
Drop Both Rush Hour and Major Interstates						
Weighted Average	0.0445	0.0270	-0.0571	0.147	0.108	-0.195
	(0.00363)	(0.00358)	(0.00377)	(0.00549)	(0.00437)	(0.00558)
N	60152	60086	71979	60152	60086	71979

Standard errors in parentheses, cluster at officer level

Rush Hour: 6am-10am and 4pm-8pm

Major Interstates: I-84, I-90, I-91, I-95, I-93, I-190, I-195, I-290, I-395, I-495

Stop observations are at cell level: town × route × year × month × day of month × hour × officer

Table A-8: Crashes / Accidents and All Traffic Stops

	Stops - Telemetric Benchmark			Stops - Accident		
	(1) Black	(2) Hispanic	(3) White	(4) Black	(5) Hispanic	(6) White
Matched at town × route × year × month × day of month × hour (2% Stops)						
Weighted Average	0.0287	0.0603	-0.0658	0.0348	0.0491	-0.0636
	(0.00562)	(0.00589)	(0.00567)	(0.00545)	(0.00645)	(0.00667)
N	5519	5963	7232	5519	5963	7232
Matched at town × route × year × month × day of month × hour bin (6% Stops)						
Weighted Average	0.0391	0.0623	-0.0731	0.0358	0.0575	-0.0687
	(0.00467)	(0.00465)	(0.00462)	(0.00504)	(0.00482)	(0.00508)
N	15440	16418	20488	15440	16418	20488
Matched at town × route × year × month × day of week × hour bin (19% Stops)						
Weighted Average	0.0484	0.0682	-0.0831	0.0433	0.0680	-0.0765
	(0.00365)	(0.00417)	(0.00419)	(0.00368)	(0.00408)	(0.00423)
N	46231	48697	61429	46231	48697	61429
Matched at town × route × year × month × day of week × shift (27% Stops)						
Weighted Average	0.0528	0.0680	-0.0853	0.0522	0.0680	-0.0822
	(0.00341)	(0.00390)	(0.00398)	(0.00380)	(0.00395)	(0.00423)
N	64571	67692	85228	64571	67692	85228
Matched at town × route × year × month × shift (69% Stops)						
Weighted Average	0.0515	0.0603	-0.0800	0.0548	0.0626	-0.0809
	(0.00264)	(0.00314)	(0.00322)	(0.00289)	(0.00317)	(0.00337)
N	175886	181552	220265	175886	181552	220265
Matched at town × route × shift (93% Stops)						
Weighted Average	0.0448	0.0522	-0.0735	0.0514	0.0551	-0.0747
	(0.00276)	(0.00285)	(0.00308)	(0.00277)	(0.00281)	(0.00323)
N	249332	255543	302081	249332	255543	302081

Standard errors in parentheses, cluster at officer level

Hour bin: 2am-5am, 6am-10am, 11am-3pm, 4pm-8pm, 9pm-1am

Shift: 7am-2pm, 3pm-10pm, 11pm-6am

Stop observations are at cell level: town × route × year × month × day of month × hour × officer

Table A-9: Crashes / Accidents and Moving Violation Stops

	Stops - Telemetric Benchmark			Stops - Accident		
	(1) Black	(2) Hispanic	(3) White	(4) Black	(5) Hispanic	(6) White
Matched at town × route × year × month × day of month × hour (3% Stops)						
Weighted Average	0.0160	0.0451	-0.0478	0.0248	0.0402	-0.0532
	(0.00594)	(0.00656)	(0.00613)	(0.00569)	(0.00785)	(0.00837)
N	3925	4197	5002	3925	4197	5002
Matched at town × route × year × month × day of month × hour bin (7% Stops)						
Weighted Average	0.0303	0.0447	-0.0562	0.0286	0.0432	-0.0560
	(0.00535)	(0.00430)	(0.00444)	(0.00622)	(0.00509)	(0.00569)
N	9437	9867	12120	9437	9867	12120
Matched at town × route × year × month × day of week × hour bin (20% Stops)						
Weighted Average	0.0420	0.0512	-0.0683	0.0382	0.0534	-0.0641
	(0.00384)	(0.00319)	(0.00339)	(0.00395)	(0.00366)	(0.00385)
N	26182	27053	33724	26182	27053	33724
Matched at town × route × year × month × day of week × shift (28% Stops)						
Weighted Average	0.0457	0.0532	-0.0717	0.0453	0.0512	-0.0669
	(0.00335)	(0.00301)	(0.00321)	(0.00395)	(0.00351)	(0.00392)
N	36012	37235	46304	36012	37235	46304
Matched at town × route × year × month × shift (72% Stops)						
Weighted Average	0.0465	0.0482	-0.0688	0.0521	0.0504	-0.0711
	(0.00239)	(0.00268)	(0.00273)	(0.00272)	(0.00286)	(0.00304)
N	99145	100740	120965	99145	100740	120965
Matched at town × route × shift (97% Stops)						
Weighted Average	0.0412	0.0430	-0.0637	0.0483	0.0453	-0.0661
	(0.00303)	(0.00259)	(0.00293)	(0.00304)	(0.00255)	(0.00328)
N	138586	139868	163625	138586	139868	163625

Standard errors in parentheses, cluster at officer level

Hour bin: 2am-5am, 6am-10am, 11am-3pm, 4pm-8pm, 9pm-1am

Shift: 7am-2pm, 3pm-10pm, 11pm-6am

Stop observations are at cell level: town × route × year × month × day of month × hour × officer

Table A-10: First Stage Regression using Inter Twilight Window of Accident

	Black		Hispanic		White	
	(1) Sample 1	(2) Sample 2	(3) Sample 1	(4) Sample 2	(5) Sample 1	(6) Sample 2
Sample 2	0.299***		0.243***		0.304***	
	(0.0336)		(0.0365)		(0.0293)	
Sample 1		0.290***		0.244***		0.299***
		(0.0332)		(0.0312)		(0.0246)
N	20021	20021	21040	21040	24246	24246
F	79.12785	76.44928	44.12516	61.52714	107.6917	147.7365
p	4.10e-18	1.41e-17	5.80e-11	1.45e-14	9.15e-24	2.49e-31

Standard errors in parentheses, cluster at town × route × shift level
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A-11: Second Stage Regression using Inter Twilight Window of Accident

	Black		Hispanic		White	
	(1) Accident	(2) Accident	(3) Accident	(4) Accident	(5) Accident	(6) Accident
Sample 1	1.899**		1.645*		1.884***	
	(0.699)		(0.669)		(0.401)	
Sample 2		2.889***		1.809**		1.738***
		(0.796)		(0.628)		(0.452)
FE Mean	0.0269	-0.0402	0.0072	-0.0130	-0.8660	-0.7443
N	20021	20021	21040	21040	24246	24246
F	7.374203	13.1813	6.048605	8.295511	22.08555	14.78723
p	.0067649	.0003015	.0141347	.0040839	3.07e-06	.0001298

Standard errors in parentheses, cluster at town × route × shift level
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A-12: Veil of Darkness without Holidays

	Veil of Darkness			Stops - Telemetric Benchmark		
	(1) Black	(2) Hispanic	(3) White	(4) Black	(5) Hispanic	(6) White
Panel A: Reweight Written Moving Violation and Verbal Stops						
disparity	0.0213***	0.0150*	-0.0226***	0.0440***	0.0379***	-0.0672***
	(0.00628)	(0.00651)	(0.00684)	(0.00433)	(0.00387)	(0.00464)
Mean of Stops	0.1816	0.2119	0.6707	0.1816	0.2119	0.6707
town × route × hour	Y	Y	Y			
officer	Y	Y	Y			
N	34133	35437	41374	34133	35437	41374
Panel B: Reweight Written Moving Violation Stops						
disparity	0.0311***	0.0216*	-0.0354***	0.0349***	0.0315***	-0.0572***
	(0.00756)	(0.00874)	(0.00872)	(0.00433)	(0.00442)	(0.00508)
Mean of Stops	0.1627	0.2019	0.6908	0.1627	0.2019	0.6908
town × route × hour	Y	Y	Y			
officer	Y	Y	Y			
N	20729	21682	24807	20729	21682	24807

Standard errors in parentheses, cluster at town × route × hour level

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Holidays are dates from Thanksgiving to Jan 1st and June 16 to end of August

Telemetric Benchmark utilizes an IV constructed from accidents occurring within the inter-twilight window

Table A-13: Veil of Darkness of Telemetric Benchmark

	(1) Black	(2) Hispanic	(3) White
Panel A: Reweight Written Moving Violation and Verbal Stops			
disparity	0.0002	0.0005	-0.0012
	(0.00045)	(0.00058)	(0.00066)
Mean of Stops	0.1377	0.1739	0.7375
town × route × hour	Y	Y	Y
officer	Y	Y	Y
N	50738	52951	61596
Panel B: Reweight Written Moving Violation Stops			
disparity	0.0012*	0.0005	-0.0011
	(0.00050)	(0.00067)	(0.00078)
Mean of Stops	0.1296	0.1708	0.7459
town × route × hour	Y	Y	Y
officer	Y	Y	Y
N	31413	32892	37594

Standard errors in parentheses, cluster at town × route × hour level
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Telemetric Benchmark utilizes an IV constructed from accidents occurring within the inter-twilight window

Table A-14: Veil of Darkness of Telemetric Benchmark without Holidays

	(1) Black	(2) Hispanic	(3) White
Panel A: Reweight Written Moving Violation and Verbal Stops and Drop Holidays			
disparity	0.0008	0.0014	-0.0024**
	(0.00056)	(0.00075)	(0.00081)
Mean of Stops	0.1377	0.1741	0.7378
town × route × hour	Y	Y	Y
officer	Y	Y	Y
N	34133	35437	41374
Panel B: Reweight Written Moving Violation Stops and Drop Holidays			
disparity	0.0013	0.0010	-0.0015
	(0.00067)	(0.00086)	(0.00097)
Mean of Stops	0.1283	0.1707	0.7481
town × route × hour	Y	Y	Y
officer	Y	Y	Y
N	20729	21682	24807

Standard errors in parentheses, cluster at town × route × hour level
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Holidays are dates from Thanksgiving to Jan 1st and June 16 to end of August
Telemetric Benchmark utilizes an IV constructed from accidents occurring within the inter-twilight window

Table A-15: Veil of Darkness of Warning Share

	(1) Black	(2) Hispanic	(3) White
Panel A: Reweight Written Moving Violation and Verbal Stops			
disparity	-0.0487***	-0.0554***	-0.0450***
	(0.01259)	(0.01133)	(0.00615)
Mean of Stops	0.7129	0.6581	0.6984
town × route × hour	Y	Y	Y
officer	Y	Y	Y
N	9744	12387	41694
Panel B: Reweight Written Moving Violation Stops			
disparity	-0.0595**	-0.0477**	-0.0376***
	(0.02140)	(0.01748)	(0.00766)
Mean of Stops	0.6593	0.6189	0.6827
town × route × hour	Y	Y	Y
officer	Y	Y	Y
N	4845	6656	26522

Standard errors in parentheses, cluster at town × route × hour level
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Holidays are dates from Thanksgiving to Jan 1st and June 16 to end of August

Table A-16: Veil of Darkness of Warning Share without Holidays

	(1) Black	(2) Hispanic	(3) White
Panel A: Reweight Written Moving Violation and Verbal Stops and Drop Holidays			
disparity	-0.0609***	-0.0547***	-0.0422***
	(0.01686)	(0.01647)	(0.00816)
Mean of Stops	0.7227	0.6618	0.7005
town × route × hour	Y	Y	Y
officer	Y	Y	Y
N	6333	7874	27940
Panel B: Reweight Written Moving Violation Stops and Drop Holidays			
disparity	-0.0924**	-0.0642*	-0.0320**
	(0.03148)	(0.02707)	(0.01011)
Mean of Stops	0.6650	0.6194	0.6842
town × route × hour	Y	Y	Y
officer	Y	Y	Y
N	2852	4021	17489

Standard errors in parentheses, cluster at town × route × hour level
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Holidays are dates from Thanksgiving to Jan 1st and June 16 to end of August

Table A-17: Veil of Darkness of Speeding Share

	(1) Black	(2) Hispanic	(3) White
Panel A: Reweight Written Moving Violation and Verbal Stops			
disparity	0.0362**	0.0623***	0.0443***
	(0.01109)	(0.01020)	(0.00543)
Mean of Stops	0.6172	0.5817	0.6567
town × route × hour	Y	Y	Y
officer	Y	Y	Y
N	9744	12387	41694
Panel B: Reweight Written Moving Violation Stops			
disparity	0.0249**	0.0182*	0.0151***
	(0.00957)	(0.00863)	(0.00316)
Mean of Stops	0.7306	0.7271	0.7392
town × route × hour	Y	Y	Y
officer	Y	Y	Y
N	4845	6656	26522

Standard errors in parentheses, cluster at town × route × hour level
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Holidays are dates from Thanksgiving to Jan 1st and June 16 to end of August

Table A-18: Veil of Darkness of Speeding Share without Holidays

	(1) Black	(2) Hispanic	(3) White
Panel A: Reweight Written Moving Violation and Verbal Stops and Drop Holidays			
disparity	0.0513***	0.0863***	0.0423***
	(0.01534)	(0.01334)	(0.00719)
Mean of Stops	0.6194	0.5844	0.6575
town × route × hour	Y	Y	Y
officer	Y	Y	Y
N	6333	7874	27940
Panel B: Reweight Written Moving Violation Stops and Drop Holidays			
disparity	0.0336*	0.0270*	0.0144***
	(0.01384)	(0.01270)	(0.00410)
Mean of Stops	0.7281	0.7307	0.7374
town × route × hour	Y	Y	Y
officer	Y	Y	Y
N	2852	4021	17489

Standard errors in parentheses, cluster at town × route × hour level
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Holidays are dates from Thanksgiving to Jan 1st and June 16 to end of August