Measuring Disparate Impact in Police Stops: Evidence from Telemetric Mobility Data

Xiaofeng Gong[1] Yuling Han[2] Susan Parker[3] **Matthew B. Ross[2]** Stephen L. Ross[1]

[1]University of Connecticut

[2]Northeastern University

[3]Northwestern University

February 6, 2026

Summary of Today's Paper: TL;DR

- **1. The Problem: Statutory Mandates**
  - 17 states mandate evaluating enforcement relative to the underlying population.
  - We solve the "Missing Denominator" problem with a localized roadway benchmark.
- **2. The Method: Calibrated Telemetry**
  - Dynamic measure of driving population using mobile GPS (MA, 2021-22).
  - **Key Contribution:** Calibrated with Crash Data (IV) to correct for device selection bias; raw data significantly understates minority presence.
- **3. Main Findings**
  - **Clear Disparate Impact:** Black & Hispanic motorists are stopped at rates significantly exceeding their roadway presence (7.1pp disparity).
  - **Validation:** Community Standard (Census) overestimates disparities (7-13pp); VOD underestimates them (20-30%) by missing structural factors.
- **4. Policy Implication**
  - Telemetry is the "Gold Standard," but Aggregated Crash Data (covering 92% of stops) is a highly accurate, low-cost alternative for government monitoring.

Scale and Social Costs of Traffic Enforcement

- 94% believe policing needs changes (Gallup, 2020)
- 21+ million traffic stops annually and have a disproportionate impact on minorities (Pierson et al. 2020)
  - Can escalate into deadly encounters (Levenson 2021; Tapp and Davis 2022)
  - Disparities in policing generate deep economic costs (Mello, 2025)
  - Erodes public trust and cooperation (Ang et al., 2021)

Legislative Response

- **Federal:** Title VI enforcement focuses on discrimination *effects*, not just intent.
- **State:** 33 states have statutes addressing profiling. 17 explicitly mandate statistical monitoring (e.g., CA's RIPA, CT's Alvin Penn Act).
- **Local:** Major cities have implicitly/explicitly prohibited pretext stops

The Methodological Hurdle: The Missing Denominator

The Policy Mandate

Statutory law increasingly requires evaluating enforcement relative to the **underlying driving population** (i.e., a "localized" version of disparate impact that conditions on time and location).

- **Federal:** DOJ Title VI prohibits practices with discriminatory effects.
- **State:** 17 states (e.g., CA's RIPA, CT's Alvin Penn Act, IL, MA, and TX) explicitly mandate statistical monitoring of the "denominator."

Defining the Correct Estimand:

- **Localized Standard:** A legal requirement to assess who is on the road *at the specific time and location of stops*.
- **vs. Broader Impact:** Broader definitions include "upstream" decisions regarding where and when police are deployed in the first place.
- **vs. Disparate Treatment:** No attempt to condition on an "at risk" population.

The Research Dilemma: Precision vs. Relevance

Researchers and policymakers have historically faced a difficult choice between two flawed approaches:

1. Rigorous but for disparate treatment

- *Methods:* Veil of Darkness (Gogger & Ridgeway 2006), Search Hit Rates (Knowles et al. 2001, Anwar & Fang 2006; Feigenberg & Miller 2022), speed discounting (Goncalves & Mello 2021)

2. Disparate Impact but with Limitations

- *Methods:* The "Community Standard" (Census Data) (McDevitt 2001) or crash benchmarks (Alpert et al. 2004).

- Rigorous methods focus on disparate treatment not impact and many are not about disparities in the decision to stop.
- Broader methods rely on largely untested and often unrealistic assumptions.

The Telemetric Solution & Its Limitations

High-frequency mobility data offers a powerful, exogenous measure of presence.

Foundational Work

- Cai et al. (2022): Compared speeding enforcement to speeding behavior.
- Aggarwal et al. (2025): Used Lyft GPS data to analyze racial disparities in speeding tickets.
- Xu et al. (2024): Analyzed stop disparities using racial composition during a representative Thursday and so cannot account for variation over time of day, day of week or season

Remaining Limitations

- **Generalizability:** Prior studies relied on specific samples (rideshare data or very restricted time windows.
- **Selection Bias:** Raw telemetric data under-represents low-income and minority households (Li et al., 2024).
- **Unanswered Questions:** Utility as a policy tool or comparison to other tests

This Paper: Validated Dynamic Roadway Benchmark

We address these remaining questions by developing a dynamic measure of the driving population using telemetric mobility data:

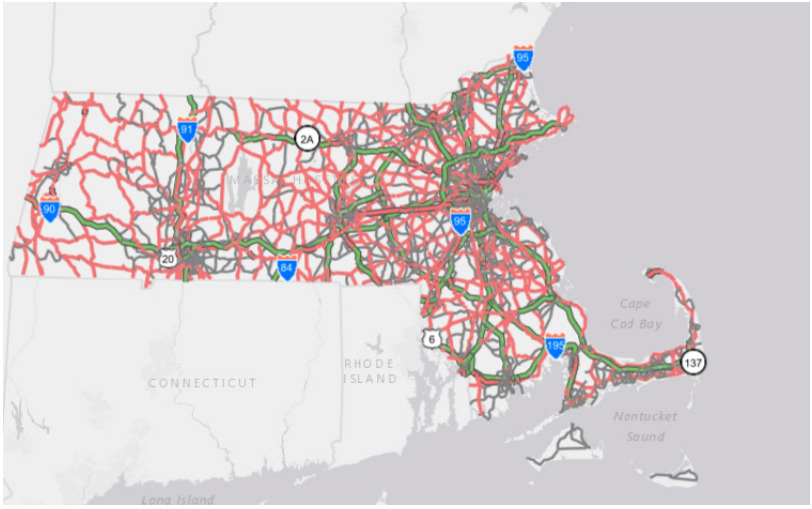
- We account for hour-by-hour and road-by-road variation in the racial composition of the roadway.
- We calibrate our race proxies and correct for selection using IV and a ground-truth measure of racial composition
- We assess disparities using data from Massachusetts State Police from 2021-22
- We validate our benchmark against spatio-temporal variation in stops as well as conventional controls used in typical models
- We use our new measure to assess several alternative tests of discrimination which are largely unvalidated, i.e. community standard, crashes, and VOD

Data Sources (Massachusetts 2021-2022)

Our analysis links three primary datasets to construct a comprehensive view of enforcement and exposure.

- **1. Police Stops (Numerator):**
  - Source: MA State Police (MASP) citations and warnings.
  - **Sample Size:** $\approx$ 416,380 stops.
  - **Scope:** Restricted to Interstates, U.S. Highways, and State Routes to match the mobility data coverage.
- **2. Telemetric Data (Denominator):**
  - Source: Anonymized mobile GPS "pings" from consistent devices.
  - **Processing:** Geocoded to roadway polygons (32ft buffer) in MA + bordering counties.
- **3. Crash Data (Calibration Target):**
  - Source: Statewide crash records identifying driver race.
  - **Role:** Acts as the "Ground Truth" for roadway presence. Since crash attendance is determined by the event (not officer discretion), it is an exogenous measure of who is on the road (Alpert et al. 2004).

Data: Analytical Sample



- We restrict to U.S. Highways and State Routes, drop secondary surface streets

Step 1: Race Proxy Construction

Goal

Predict the racial composition of motorists (R) in a specific cell c defined by Town $\times$ Highway $\times$ Date $\times$ Hour.

"Home" Assignment

- Vendor observes a device i at their "home" and provides us w/ a Census Block Group of residence.
- We assign a race likelihood (L_{ij}) to the device based on the demographics of that Block Group.
- We aggregate these likelihoods to the road-segment level (Pr_{cj}).

The Problem

This raw proxy (Pr_{cj}) suffers from **Measurement Error** and **Selection Bias** (e.g., wealth determines smartphone ownership types).

Step 2: Split-Sample IV Calibration of Race

We correct the raw proxy using a Split-Sample Instrumental Variables (IV) approach.

1. First Stage (Isolating Signal): Regress a Hold-out Proxy (Pr^H) on the Analysis Proxy (Pr^A) to purge random measurement error.

$$Pr^H_{mcdj} = \alpha_j Pr^A_{cdj} + \delta_{dj} + \mu_{mcdj} \tag{1}$$

2. Calibration (Correcting Bias): Regress the actual Race of Crash Drivers (R_{mcdj}) on the predicted proxy.

$$R_{mcdj} = \beta_j(\hat{\alpha}_j \hat{Pr}_{cdj}) + \gamma_{dj} + \epsilon_{mcdj} \tag{2}$$

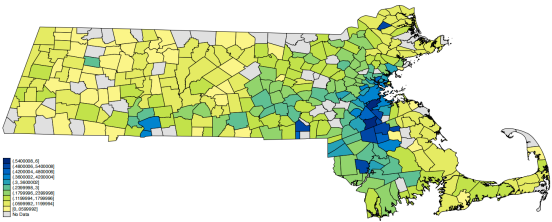
Validation 1: Ability to Explain Spatial & Temporal Variation

Does our measure predict actual changes in the racial composition of stops?

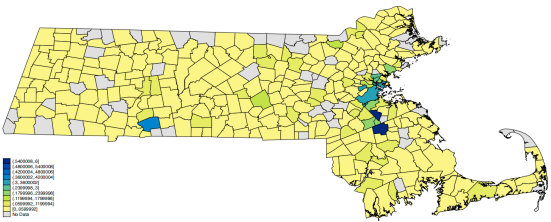
Variation Source	**Telemetric Benchmark**	**Census Benchmark**
Across Towns	$\approx$ **65% Explained**	Increases Variance
Hour of Day / Day of Week	$\approx$ **58% Explained**	0% Explained

Key Takeaway: The Telemetric measure captures the "shape" of enforcement (temporal and geographic patterns) with high fidelity, whereas the Census benchmark fails to account for dynamic population shifts.

Geographic Variation: Stops vs. Census

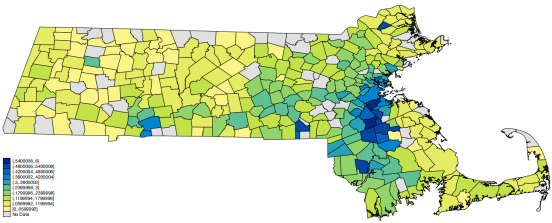


- Stop Data, Black

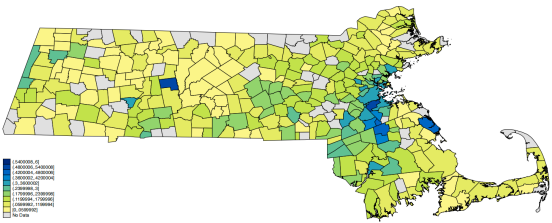


- Census Data, Black

Geographic Variation: Stops vs. Telemetric

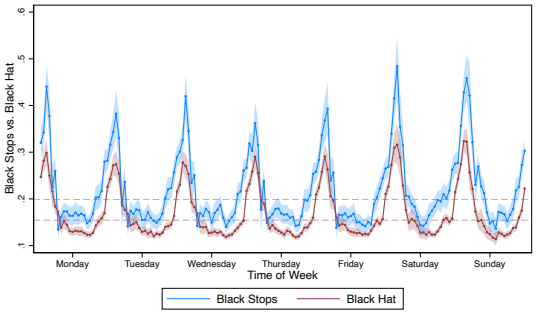


- Stop Data, Black

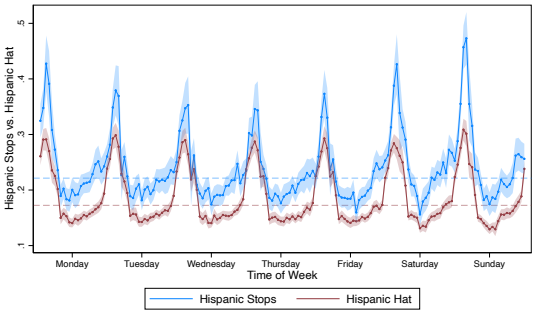


- Telemetric Data, Black

Temporal Variation: Stops vs. Telemetric



- Stops vs. Telemetric for Share Black



- Stops vs. Telemetric for Share Hispanic

Validation 2: Officer Propensities

We benchmark our measure against high-dimensional fixed-effects models frequently used in the literature (Ridgeway and MacDonald, 2014; Gong et al., 2025).

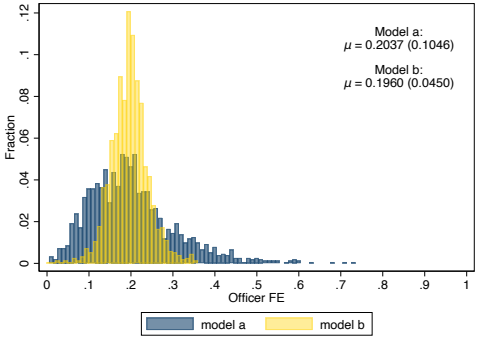
Document Across Officer Heterogeneity in Police Stops

- **Model A:** Controls: officer FE and highway by town by shift FE .
- **Model B:** Controls: officer FE.
- **The Test:** We compare officer propensity estimates (Bayesian shrunk FE's) from models A & B w/ racial composition of stops vs. racial composition of stops - telemetric racial compositions

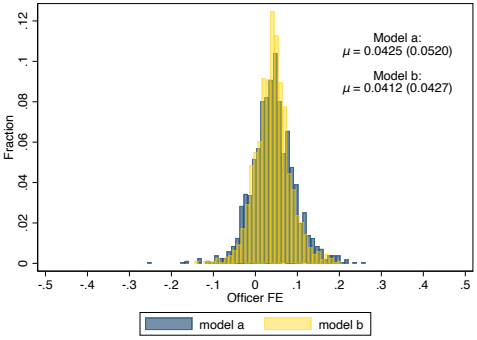
The Result

- The officer propensity estimates from both models are highly correlated ($\rho \approx 0.90$).
- **Implication:** Our measure successfully captures granular exposure risk without requiring the massive data density needed for high-dimensional fixed effects.

Officer Variation



- Stops



- Stops - Telemetric

- Figure 1 (a) and Figure 1 (b), $\rho = 0.733$
- Figure 1 (b) and Figure 2 (a), $\rho = 0.868$
- Figure 2 (a) and Figure 2 (b), $\rho = 0.873$

Result 1: Clear Evidence of Disparate Impact

The "Level Shift"

While our measure tracks the *fluctuations* of policing, there is a persistent gap in the *levels*.

Aggregate Disparities:

- **Black & Hispanic Share of Stops:** 34.75%
- **Black & Hispanic Share of Drivers:** 15.27%
- Comparable to Xu et al.'s 20 percentage point disparity using predicted distribution on a typical Thursday.

Regression Estimates (Net of Roadway Composition):

- **Black Motorists:** +4.4 percentage points (Base: 19.88%)
- **Hispanic Motorists:** +4.9 percentage points (Base: 22.14%)
- **White Motorists:** -7.1 percentage points (Base: 65.25%)

All estimates are statistically significant.

Result 2: The Failure of the "Community Standard"

We tested the traditional method of comparing stops to Residential Census Population.

The Finding

- The Community Standard **overestimates disparities** by 7 to 13 percentage points.

Why Does it Fail?

- **Commuting Patterns:** People do not drive only where they live.
- **Vehicle Ownership:** Minority households often have lower vehicle ownership rates and higher transit usage (Bunten et al., 2024).
- **Robustness:** The bias of the test persists even when we exclude rush hour (commuters) and interstates (through-traffic).
- **Empirical Observation:** We see an increase in the number of stops at night and, based on the telemetric data, minority motorists are much more present on the roadway at night.

Result 3: The Crash Benchmark Breakthrough

The Skepticism

Critics argue Crash Data is too rare (small sample size) to be a valid benchmark.

Our Solution: Aggregation

- We aggregated crash data to the **Highway** $\times$ **Town** $\times$ **Shift** level.
- This expanded coverage from 2% of stops to **92% of stops** by expanding the time periods and the dates over which an accident might be used to provide a race counterfactual..

The Finding

- The Aggregated Crash Benchmark produces disparity estimates nearly identical to our Telemetric "Gold Standard."
- **Black Disparity (Telemetric w/ 92% coverage):** 4.5% vs **(Crash w/ 92% coverage):** 5.1%

Result 4: Veil of Darkness (VOD)

- **Method:** Exploits darkness to mask race, theoretically isolating bias (Grogger and Ridgeway, 2006).
- **Result:** VOD estimates are consistently **smaller** ($\approx$ 20-30%) than telemetric estimates.

Interpretation

- VOD captures **Disparate Treatment** (bias dependent on visibility).
- Telemetry captures **Disparate Impact** (bias + deployment + structural factors).
- *Confounder:* We find speeding enforcement increases in daylight for *all* groups, complicating VOD assumptions.

Conclusions & Policy Takeaways

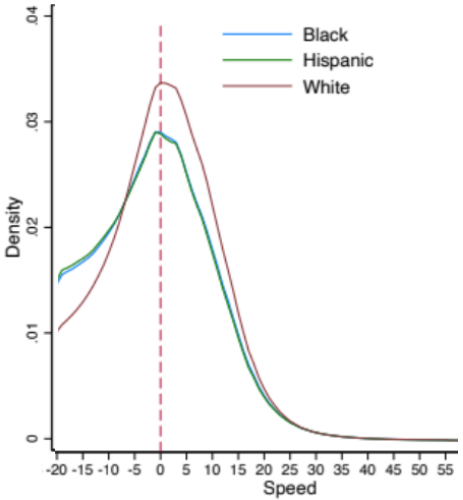
1. **Measurement Matters:** The "Community Standard" is scientifically invalid and produces inflated disparity estimates.
2. **Clear Disparate Impact:** We document substantial over-representation of minority motorists in MA using the most rigorous counterfactual to date.
3. **A Feasible Solution:**
   - Telemetry is the "Gold Standard," but is expensive.
   - **Aggregated Crash Data** is a highly accurate, zero-cost alternative for government monitoring.

Future Work w/ Telemetric Data: Disparate Impact Paper

Disparate Impact vs. Disparate Treatment: Use data from MA and TX to theoretically and conceptually explore.

- If Black/Hispanic vs. White motorists drive at different speeds, decision about what speed to stop matters
- How much of the observed disparity can be decomposed into disparate treatment vs. disparate impact?
- Officers vary in their "stop threshold"- how much of the across officer disparities can be decomposed this way?

Future Work w/ Telemetric Data: Disparate Impact Paper



- Speed distribution of motorists must be different for there to be disparate impact of "stop threshold"

References I

Pradhi Aggarwal, Alec Brandon, Ariel Goldszmidt, Justin Holz, John A. List, Ian Muir, Gregory Sun, and Thomas Yu. High-frequency location data show that race affects citations and fines for speeding. *Science*, 387(6741):1397–1401, 2025. doi: 10.1126/science.adp5357. URL `https://www.science.org/doi/abs/10.1126/science.adp5357`.

Desmond Ang, Panka Bencsik, Jesse Bruhn, and Ellora Derenoncourt. Police violence reduces civilian cooperation and engagement with law enforcement. *HKS Working Paper*, (RWP21-022), September 2021.

Devin Michelle Bunten, Ellen Fu, Lyndsey Rolheiser, and Christopher Severen. The problem has existed over endless years: Racialized difference in commuting, 1980–2019. *Journal of Urban Economics*, 141:103542, May 2024. doi: 10.1016/j.jue.2023.103542. URL `https://doi.org/10.1016/j.jue.2023.103542`.

References II

William Cai, Johann Gaebler, Justin Kaashoek, Lisa Pinals, Samuel Madden, and Sharad Goel. Measuring racial and ethnic disparities in traffic enforcement with large-scale telematics data. *PNAS Nexus*, 1(4):pgac144, 07 2022. ISSN 2752-6542. doi: 10.1093/pnasnexus/pgac144. URL https://doi.org/10.1093/pnasnexus/pgac144.

Xiaofeng Gong, Yuling Han, Susan T. Parker, Matthew B. Ross, and Stephen L. Ross. Are police officers monolithic in their treatment of minority motorists - how heterogeneous are officer traffic stop patterns over race? Technical report, 2025.

Jeffrey Grogger and Greg Ridgeway. Testing for racial profiling in traffic stops from behind a veil of darkness. *Journal of the American Statistical Association*, 101(475):878–887, 2006. doi: 10.1198/016214506000000168.

Zhenlong Li, Huan Ning, Fengrui Jing, and M Naser Lessani. 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, January 2024.

References III

Steve Mello. Fines and financial wellbeing. *The Review of Economic Studies*, 92(5): 3340–3374, October 2025. URL https://doi.org/10.1093/restud/rdae111.

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

Wenfei Xu, Michael Smart, Nebiyou Tilahun, Sajad Askari, Zachary Dennis, Houpu Li, and David Levinson. The racial composition of road users, traffic citations, and police stops. *Proceedings of the National Academy of Sciences*, 121(24):e2402547121, 2024. doi: 10.1073/pnas.2402547121. URL https://www.pnas.org/doi/abs/10.1073/pnas.2402547121.