A Veil-of-Darkness Test of Racial Disparity in North Carolina Traffic Stops, 2010–2015, with a Post-Stop Outcome (Search and Hit-Rate) Analysis

July 11, 2026

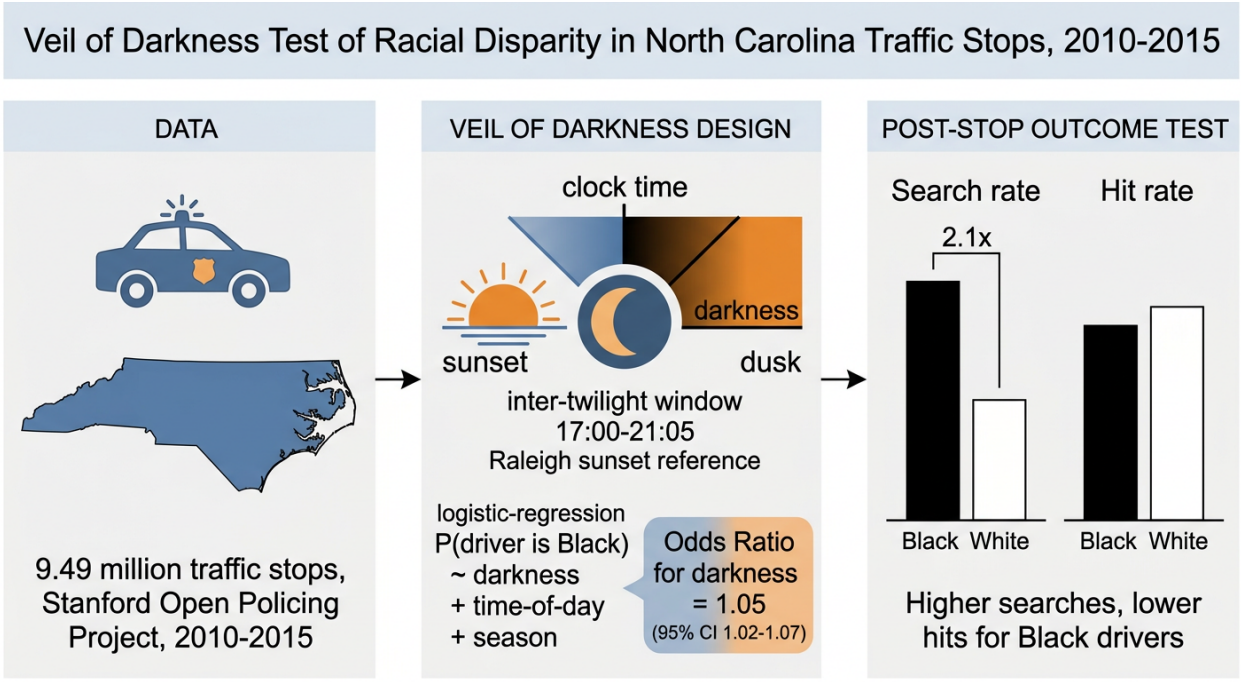

Figure 1: Graphical abstract. The analysis links 9.49 million standardized North Carolina traffic-stop records (Stanford Open Policing Project, 2010–2015) to daily solar geometry for Raleigh, NC. Within the *inter-twilight window* (17:00–21:05 local clock time), a logistic regression estimates the odds that a stopped driver is Black as a function of a darkness indicator, net of time-of-day and season. A complementary post-stop *outcome test* compares search rates and contraband hit rates by driver race.

Abstract

The "veil of darkness" (VOD) test exploits the fact that officers can less readily perceive a driver's race after dark to probe whether the racial composition of traffic stops depends on visibility, holding clock time fixed [9]. We apply the VOD test to the Stanford Open Policing Project's standardized North Carolina statewide file, restricted to 2010–2015 (9,490,673 stops). Using the `astral` library, we compute daily sunset and civil dusk for a single statewide reference point (Raleigh, NC; 35.7796°N, 78.6382°W) and define the inter-twilight window as the clock-time interval between the earliest sunset (17:00:27) and the latest civil dusk (21:05:08) observed over the period. Within this window we fit logistic regressions for the probability that a stopped driver is Black (versus White) on a darkness indicator, controlling for 30-minute time-of-day bins, season, and year, with county-clustered standard errors. Contrary to the classic VOD prediction

of an odds ratio below one, the fully adjusted darkness odds ratio is OR = 1.045 (95% CI 1.024–1.067; $p = 2.1 \times 10^{-5}$): the odds a stopped driver is Black are about 4.5% *higher* in darkness than daylight once clock time is held fixed. Adding time-of-day controls shrinks the raw odds ratio from 1.121 to 1.045, showing that most of the unconditional day/night gap is confounded with clock time. In a complementary outcome test, Black drivers are searched at 2.10× the White rate (3.67% vs. 1.75%), a disparity robust to year and county fixed effects (adjusted OR = 2.019), while unadjusted contraband hit rates are *lower* for Black drivers (29.6% vs. 31.4%). The joint pattern of higher searches and lower hits is the textbook outcome-test signature of a lower evidentiary search threshold, although the hit-rate gap becomes statistically indistinguishable from zero once county and year are controlled (adjusted OR = 0.986, $p = 0.19$), suggesting the raw hit-rate gap is largely a geographic (compositional) artifact. We release runnable code, the analysis dataset, and all model-output CSVs, and we interpret the estimates under the standard caveats of the VOD design and the infra-marginality problem of outcome tests.

1 Introduction and Background

Racial disparities in the enforcement of traffic law are among the most extensively studied and contested questions in the empirical social sciences. Descriptively, Black drivers in the United States are stopped and searched at markedly higher rates than White drivers, a pattern documented across jurisdictions and confirmed at national scale by analyses of tens of millions of records [3, 18, 24]. The central analytical difficulty is not measuring disparity but attributing it: an observed excess of minority stops relative to a population baseline may reflect discriminatory enforcement, but it may equally reflect differences in driving exposure, vehicle characteristics, residential geography, or the spatial deployment of patrol resources. This is the "denominator problem" or benchmarking problem: to interpret a stop rate one must know the racial composition of the population genuinely at risk of a stop, which is almost never observed [5, 19].

The veil-of-darkness (VOD) test, introduced by Grogger and Ridgeway [9], is a quasi-experimental design that sidesteps the benchmarking problem by conditioning on visibility rather than on an external population count. Its logic is disarmingly simple. If officers engage in racial profiling at the stop decision, they must be able to perceive the driver's race before deciding to stop; after dark, race is substantially harder to discern. Consequently, if profiling operates through visual cues, the share of stopped drivers who are Black should *fall* when the same stretch of road is dark rather than light. The design controls for the confounding influence of clock time, the strongest determinant of both who is driving and how police are deployed, by restricting attention to the *inter-twilight window*: the band of clock times that is sometimes in daylight and sometimes in darkness across the year, because sunset drifts by several hours between the winter and summer solstices and shifts abruptly at daylight-saving transitions. Comparing stops at, say, 7:30 p.m. in June (light) with 7:30 p.m. in December (dark) holds clock time fixed while varying only illumination, approximating a natural experiment in the visibility of race.

Formally, the VOD test estimates the odds that a stopped driver belongs to the group of interest as a function of a binary darkness indicator, with fixed effects for narrow time-of-day bins and for season. Under the identifying assumption that the racial composition of drivers on the road at a given clock time and season does not itself change between light and dark, an odds ratio below one is interpreted as evidence consistent with profiling against the group, and an odds ratio near one as the absence of visibility-mediated differential treatment. The method has become a workhorse of the field: it underpins the Stanford Open Policing Project's national analysis [18], is embedded in official state reporting, and has been extended and stress-tested by a substantial methodological literature [11, 19].

The VOD test speaks only to the *stop* decision. A second, older tradition of discrimination testing addresses *post-stop* decisions, particularly the decision to search. The "outcome test," rooted in Becker's economics of discrimination [4] and formalized for policing by Knowles et al. [14], reasons about the success rate of searches. If officers search to maximize the probability of finding contraband and apply a common evidentiary threshold to all drivers, the marginal search of every group should yield contraband at the same rate; a group searched down to a *lower* threshold will exhibit a *lower* hit rate. The same rational-choice framework connects the fairness of the search rule to its crime-control effectiveness, so departures from equal hit rates carry both equity and efficiency implications [16]. Thus the diagnostic signature of a discriminatory (lower) threshold applied to a group is the joint occurrence of a *higher* search rate and a *lower* hit rate. This logic has been applied widely, from Maryland interstate searches [14] to New York City's stop-and-frisk program [7, 8] and to national use-of-force data [6].

Both tests carry well-understood caveats. The VOD test can be biased if driving behavior is itself endogenous to darkness in a race-correlated way, or if the darkness indicator is measured with error because a single reference location cannot capture longitudinal variation in sunset across a wide state [11]. The outcome test suffers from the problem of *infra-marginality*: because hit rates average over the entire searched population rather than isolating the marginal search, unequal hit rates can arise from differing risk distributions even absent a threshold difference, and equal hit rates need not imply an equal threshold [2, 17, 21]. Subsequent work has proposed alternative tests that relax the strong distributional assumptions of the original outcome test, for example by exploiting variation across officers or over time [1]. These limitations motivate reporting the two tests jointly, with controls, and interpreting them as convergent but not dispositive evidence.

This report applies both tests to North Carolina, a state whose mandatory stop-data collection statute has made it a canonical setting for this literature [3, 24]. We restrict the Stanford Open Policing Project's standardized statewide file to 2010–2015 and pursue four objectives: (i) construct a defensible inter-twilight window from local solar geometry and document the sunset-time source; (ii) estimate the darkness odds ratio for the probability a stopped driver is Black under a graded sequence of controls; (iii) conduct an outcome test of search and hit rates by race, with formal significance tests; and (iv) assess how county and year controls reshape the raw disparities. All code, the analytical dataset, and the model-output CSVs are released to make the analysis fully reproducible.

2 Data and Methods

2.1 Data source and sample restriction

The analysis uses the Stanford Open Policing Project (OPP) standardized North Carolina statewide file, release `2020_04_01` (dataset `druid:yg821jf8611`) [18, 22]. The OPP pipeline harmonizes raw agency records into a common schema with consistent field names and race categories, which is what makes cross-jurisdiction and longitudinal analysis tractable. From the raw file of 20,286,645 stops we retained all records with a stop date between 1 January 2010 and 31 December 2015 inclusive, yielding **9,490,673 stops** (46.78% of the raw file). The retained records are distributed smoothly across years, declining gently from 1,723,939 stops in 2010 to 1,422,319 in 2015.

Table 1 summarizes the key fields. Driver race (`subject_race`) is complete, with White (5,391,170) and Black (3,065,082) drivers together accounting for 89.1% of stops; Hispanic, Asian/Pacific Islander, other, and unknown drivers make up the remainder. Following the standard convention

Table 1: Key fields in the filtered 2010–2015 North Carolina sample (N = 9,490,673 stops).

Field	Description	Value / count
Stops, 2010–2015	After date filter (from 20.29M raw)	9,490,673
White drivers	`subject_race == white`	5,391,170
Black drivers	`subject_race == black`	3,065,082
Missing stop time	`time` field absent	4,669,497 (49.20%)
Searches conducted	`search_conducted == TRUE`	231,393
Contraband recovered	`contraband_found == TRUE`	68,101

in the VOD and outcome-test literatures, all inferential comparisons are restricted to Black versus White drivers [9, 14]. Two data limitations are material and are handled explicitly. First, the `time` field is missing for 4,669,497 stops (49.20%); because the VOD test is intrinsically time-based, these records cannot enter the inter-twilight sample, and we return to the selection implications in Section 5. Second, `contraband_found` is defined only for the 231,393 stops in which a search was conducted (it is missing by design for the remaining 97.56%), so hit-rate calculations are conditioned on search occurrence.

2.2 Reference location and sunset/dusk computation

Solar geometry was computed with the `astral` Python library (v3.2) [13], which implements standard astronomical algorithms for the position of the sun. Because the OPP file does not geolocate individual stops finely enough to assign each a local horizon, and to keep the darkness classification transparent and reproducible, we adopt a *single statewide reference location*: Raleigh, North Carolina (latitude 35.7796°N, longitude 78.6382°W), the state capital and an approximate population centroid. For every one of the 2,191 days in 2010–2015 we computed two solar events in the `America/New_York` timezone:

- **Sunset**, the instant the sun's upper limb reaches the horizon; and
- **Civil dusk** (end of civil twilight), the instant the sun reaches 6° below the horizon, after which artificial lighting is generally required and race becomes hard to perceive.

Times are timezone-aware, and the `zoneinfo` database handles the EST/EDT daylight-saving transitions automatically; we verified that the local UTC offset flips from −05:00 to −04:00 on 8 March 2015 and back on 1 November 2015. The full daily table (sunset, dusk, and UTC offset for all 2,191 days) is released as `raleigh_sun_2010_2015.csv` to document provenance.

2.3 Inter-twilight window and the darkness indicator

The inter-twilight window is the clock-time band that experiences both daylight and darkness over the study period. We define its bounds from the extreme solar events observed across 2010–2015: the lower bound is the *earliest sunset* (17:00:27, on 5 December 2010) and the upper bound is the *latest civil dusk* (21:05:08, on 27 June 2010). A stop is included in the VOD sample if its local clock time t satisfies $17{:}00{:}27 \le t \le 21{:}05{:}08$, a window 244.7 minutes wide. Restricting to this band ensures that every included clock time is genuinely ambiguous—sometimes light, sometimes dark—so that the darkness contrast is not mechanically confounded with the level of clock time.

Within the window, each stop is classified as *dark* if it occurred at or after that calendar date's local sunset, and *light* otherwise:

$$\texttt{darkness}_i = \mathbb{1}[\, t_i \geq \text{sunset}(\text{date}_i)\,]\,. \tag{1}$$

We use sunset (rather than dusk) as the darkness threshold for the main specification because it is the sharper, more widely reported boundary; the brief civil-twilight period between sunset and dusk is retained inside the window as part of the ambiguous band. Of the 9,490,673 filtered stops, 4,669,497 lack a time and are dropped, 4,821,176 have a valid time, and **824,776** fall inside the inter-twilight window. Restricting to Black and White drivers leaves **729,785** stops (315,834 Black; 413,951 White) for the regression.

2.4 Veil-of-darkness regression specification

Let $y_i = 1$ if stop i involves a Black driver and $y_i = 0$ if White. We model the probability $\pi_i = \Pr(y_i = 1)$ with logistic regression,

$$\text{logit}(\pi_i) = \alpha + \beta\, \texttt{darkness}_i + \gamma' \mathbf{b}_i + \delta' \mathbf{s}_i + \theta' \mathbf{y}_i, \tag{2}$$

where $\mathbf{b}_i$ are indicator variables for 30-minute clock-time bins (`tod_bin`), $\mathbf{s}_i$ are season indicators (Winter/Spring/Summer/Fall), and $\mathbf{y}_i$ are year indicators. The quantity of interest is the darkness odds ratio $\text{OR} = e^{\beta}$, the multiplicative change in the odds that a stopped driver is Black when the stop occurs in darkness rather than daylight at the same clock time and season. We report four nested specifications: (1) raw (`darkness` only); (2) adding time-of-day bins and season; (3) adding year; and (4) Model 3 re-estimated with cluster-robust standard errors clustered on `county_name` (100 counties) to accommodate spatial correlation in enforcement practice. Models were fit as binomial GLMs with a logit link in `statsmodels` [20]; the 95% confidence interval for the odds ratio is $\exp(\hat{\beta} \pm 1.96\, \widehat{\text{SE}}(\hat{\beta}))$.

2.5 Post-stop outcome test

The outcome test compares two rates by race. The *search rate* is the fraction of stops that lead to a search, and the *hit rate* is the fraction of searches that recover contraband [4, 14]. We compute both rates for Black and White drivers on two samples: the full 2010–2015 dataset (8,456,252 Black/White stops) and the inter-twilight sample (729,785 stops), the latter to check that the VOD subsample is not aberrant. Each rate is reported with a Wilson 95% confidence interval. Group differences are tested with a chi-squared test of independence (no continuity correction) and a two-proportion z-test, and summarized by the disparity (relative-risk) ratio, the Black rate divided by the White rate, with a log-method 95% interval. As a robustness check we fit controlled logistic regressions on the full dataset with the driver-race indicator plus year and county fixed effects: for the search decision, $\texttt{search} \sim \texttt{is\_black} + C(\texttt{year}) + C(\texttt{county})$ over all stops; and for the hit outcome, $\texttt{contraband} \sim \texttt{is\_black} + C(\texttt{year}) + C(\texttt{county})$ over searched drivers only (counties with fewer than 25 searches collapsed to an "Other" category to avoid separation, and the hit model fit with a Newton solver). Numerical work used `NumPy` [10], `pandas` [15], `SciPy` [23], and `Matplotlib` [12].

2.6 Reproducibility

The pipeline is organized as four scripts (download/filter, sunset construction, VOD regression, and outcome analysis). It emits the analytical dataset `nc_stops_inter_twilight.csv.gz` (824,776

rows), the daily solar table, the VOD coefficient CSV, and the outcome-test CSVs, together with the figures reproduced here. Listing 1 shows the core of the darkness-indicator construction and the VOD model fit.

```python
from astral import LocationInfo
from astral.sun import sun
import pandas as pd, statsmodels.formula.api as smf

# --- daily sunset / civil dusk for Raleigh, NC ---
loc = LocationInfo("Raleigh", "USA", "America/New_York", 35.7796, -78.6382)
def solar(day):
    s = sun(loc.observer, date=day, tzinfo=loc.timezone, dawn_dusk_depression=6.0)
    return s["sunset"], s["dusk"] # tz-aware datetimes

# --- inter-twilight window from period extremes ---
lo = earliest_sunset # 17:00:27 (2010-12-05)
hi = latest_civil_dusk # 21:05:08 (2010-06-27)
df = df[(df.clock_time >= lo) & (df.clock_time <= hi)].copy()

# --- darkness indicator and controls ---
df["darkness"] = (df.local_dt >= df.date.map(sunset_of_date)).astype(int)
df["tod_bin"] = (df.minutes // 30).astype("category") # 30-min FE

# --- Black vs White; fit VOD logistic regression ---
d = df[df.subject_race.isin(["black", "white"])].copy()
d["is_black"] = (d.subject_race == "black").astype(int)
m = smf.logit("is_black ~ darkness + C(tod_bin) + C(season) + C(year)", d).fit()
OR = np.exp(m.params["darkness"]) # darkness odds ratio
```

Listing 1: Core of the darkness-indicator construction and the veil-of-darkness logistic regression (abridged).

3 Results: The Veil-of-Darkness Test

3.1 Unconditional day/night contrast

Before conditioning on clock time, the raw composition of the inter-twilight sample already differs by lighting. Among the 729,785 Black and White stops in the window, the probability that a stopped driver is Black is 0.4169 in daylight ($n = 317{,}445$) and 0.4450 in darkness ($n = 412{,}340$), a difference of +2.81 percentage points. Taken at face value this is the *opposite* of the classic veil prediction: the share Black rises, not falls, after dark. As the model sequence below makes clear, however, most of this raw contrast reflects the fact that darkness in the window is systematically correlated with later clock times, at which the driving population is disproportionately Black for reasons unrelated to visibility.

3.2 Adjusted darkness odds ratios

Table 2 and Figure 2 report the darkness odds ratio across the four specifications. In the raw model the odds a stopped driver is Black are OR = 1.121 times as high in darkness as in daylight (95% CI 1.111–1.132). Introducing 30-minute time-of-day bins and season (Model 2) collapses the estimate to OR = 1.046 (95% CI 1.030–1.061), absorbing roughly two-thirds of the raw gap; this shrinkage is the empirical fingerprint of clock-time confounding, exactly the confound the inter-twilight design

Table 2: Darkness odds ratios for the probability a stopped driver is Black (versus White), inter-twilight sample, $N = 729{,}785$. The odds ratio is e^{β} on the `darkness` indicator; 95% CI = $\exp(\hat{\beta} \pm 1.96\,\mathrm{SE})$.

Specification	OR (darkness)	95% CI	p-value	Covariance
Model 1: Raw	1.121	[1.111, 1.132]	4.4×10^{-127}	nonrobust
Model 2: + Time-of-day & Season	1.046	[1.030, 1.061]	3.3×10^{-9}	nonrobust
Model 3: + Year	1.045	[1.030, 1.061]	4.1×10^{-9}	nonrobust
Model 4: Model 3 + Cluster(county)	1.045	[1.024, 1.067]	2.1×10^{-5}	cluster (100)

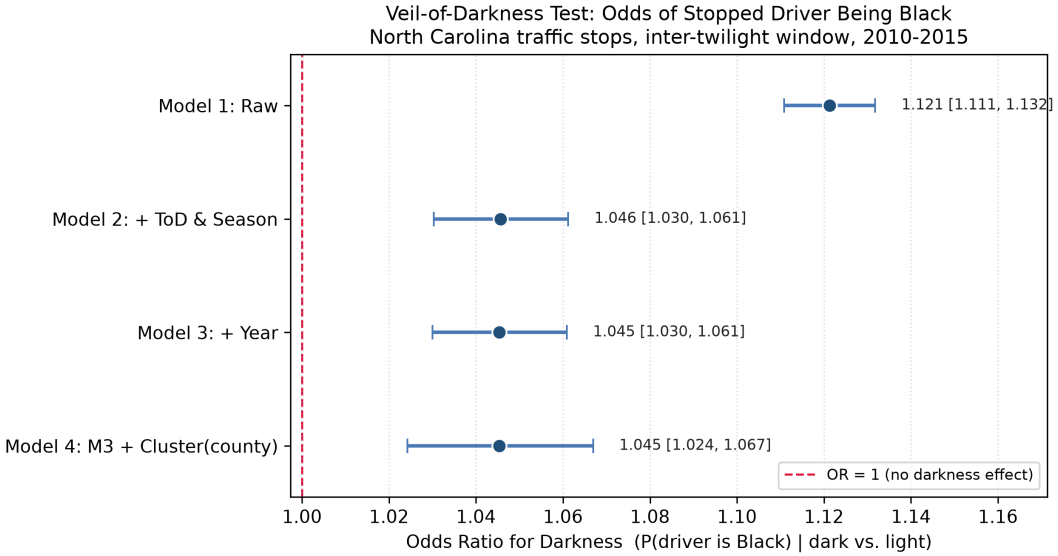

Figure 2: Veil-of-darkness odds ratios across specifications. Each point is the estimated odds ratio on the darkness indicator for the probability a stopped driver is Black; whiskers are 95% confidence intervals. The dashed line at OR = 1 marks no visibility effect. Adding 30-minute time-of-day bins and season shrinks the raw estimate from 1.12 to 1.05; the adjusted effect lies significantly *above* one, the opposite of the classic "veil" prediction.

is built to remove. Adding year fixed effects (Model 3) leaves the estimate essentially unchanged at OR = 1.045 (95% CI 1.030–1.061). Clustering standard errors on county (Model 4) holds the point estimate fixed and widens the interval to 1.024–1.067; the effect remains statistically significant ($p = 2.1 \times 10^{-5}$).

3.3 Interpretation of the darkness effect

The canonical VOD prediction is OR < 1: if profiling operates through visual perception of race, darkness should *reduce* the Black share of stops. We instead estimate OR = 1.045, significantly above one, in every adjusted specification. Three readings are consistent with this. First, and most conservatively, the adjusted estimate is close to one in magnitude—a 4.5% odds shift is small relative to the twofold search disparities documented below—so the visibility-mediated contrast provides *little evidence of profiling at the stop decision of the classic form.* Second, the large gap between the raw (1.121) and adjusted (1.045) estimates demonstrates that naive day/night comparisons are badly confounded by clock time and that the inter-twilight design materially corrects this. Third,

Table 3: Search and hit rates by driver race. Search rate = searches/stops; hit rate = contraband recovered/searches. Disparity ratio = Black rate ÷ White rate. 95% intervals are Wilson (rates) and log-method (ratios). p-values from the chi-squared test of independence.

Sample	Metric	Race	Rate (95% CI)	Num./Denom.	Disparity (B/W)	p
Full 2010–2015	Search	Black	3.674% [3.653, 3.695]	112,621 / 3,065,082	2.10 [2.08, 2.12]	$< 10^{-300}$
		White	1.748% [1.737, 1.759]	94,245 / 5,391,170		
	Hit	Black	29.64% [29.37, 29.91]	33,380 / 112,621	0.945 [0.933, 0.957]	2.0×10^{-17}
		White	31.36% [31.07, 31.66]	29,559 / 94,245		
Inter-twilight	Search	Black	5.455% [5.376, 5.535]	17,228 / 315,834	1.88 [1.84, 1.93]	$< 10^{-300}$
		White	2.895% [2.844, 2.946]	11,982 / 413,951		
	Hit	Black	30.13% [29.45, 30.82]	5,191 / 17,228	0.847 [0.820, 0.876]	2.0×10^{-22}
		White	35.55% [34.70, 36.42]	4,260 / 11,982		

an odds ratio above one is precisely the pattern that the endogenous-driving critique anticipates: if the mix of drivers on the road after dark is more heavily Black for reasons of exposure, employment schedules, or differential night-time patrol allocation, the darkness coefficient will be positive even absent any visibility-based profiling [11]. Because the VOD test identifies only a visibility-mediated contrast and not the total level of disparity, we do not interpret OR > 1 as evidence of "reverse" discrimination; we interpret it as the absence of the specific day/night signature the test is designed to detect, conditional on the design's assumptions.

4 Results: Post-Stop Outcomes by Race

The stop decision is only the first exercise of discretion; the decision to search a stopped driver is where the outcome test applies. Table 3 and Figure 3 report search and hit rates by race for both the full dataset and the inter-twilight sample.

4.1 Search rates

In the full 2010–2015 dataset, Black drivers are searched in 3.674% of stops (95% CI 3.653%–3.695%) versus 1.748% for White drivers (95% CI 1.737%–1.759%). The disparity ratio is **2.10** (95% CI 2.08–2.12): a stopped Black driver is more than twice as likely to be searched as a stopped White driver. The difference is overwhelmingly significant ($\chi^2(1) = 30{,}380$, $p < 10^{-300}$; two-proportion $z = 174.3$). The inter-twilight sample shows the same qualitative pattern with somewhat higher levels—5.455% for Black versus 2.895% for White drivers, a disparity ratio of 1.88—confirming that the VOD subsample is not anomalous with respect to search behavior.

4.2 Hit rates

The hit rate—contraband recovered per search—runs in the opposite direction. In the full dataset, searches of Black drivers recover contraband 29.639% of the time (95% CI 29.373%–29.907%), *below* the 31.364% hit rate for White drivers (95% CI 31.069%–31.661%). The disparity ratio is 0.945 (95% CI 0.933–0.957), significantly below one ($\chi^2(1) = 72.1$, $p = 2.0 \times 10^{-17}$). The gap is wider in the inter-twilight sample: 30.131% for Black versus 35.553% for White drivers, a ratio of 0.847 ($p = 2.0 \times 10^{-22}$).

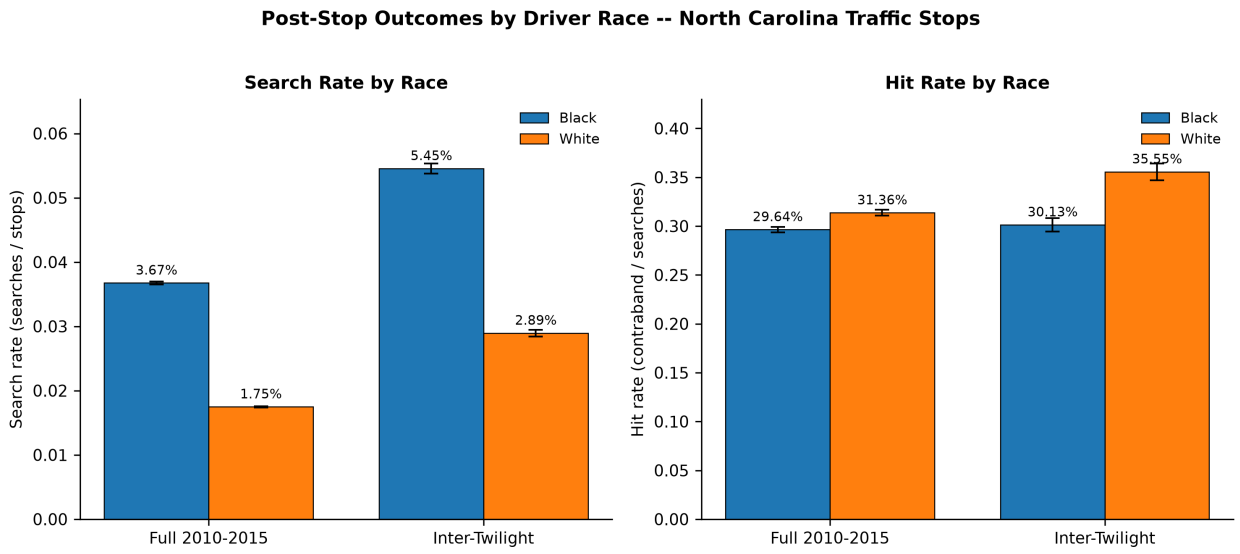

Figure 3: Post-stop outcomes by driver race. Left: search rate (searches per stop) for Black and White drivers in the full and inter-twilight samples; Black drivers are searched roughly twice as often. Right: contraband hit rate (contraband per search); Black drivers exhibit *lower* hit rates in both samples. Error bars are 95% Wilson intervals. The joint pattern—more searches, fewer hits—is the classic outcome-test signature of a lower search threshold.

4.3 The outcome-test signature

The two findings combine into the textbook diagnostic of Knowles et al. [14]: Black drivers face a *higher* search rate but a *lower* hit rate. Under the model in which officers search to maximize contraband recovery against a common threshold, this combination implies that Black drivers are searched down to a lower evidentiary bar than White drivers—statistical evidence consistent with a discriminatory search threshold. The pattern mirrors the national findings of the Stanford Open Policing Project [18] and the stop-and-frisk analyses of Gelman et al. [7] and Goel et al. [8].

4.4 Do county and year controls change the picture?

The unadjusted comparisons pool across 100 counties that differ enormously in demographics, drug-enforcement intensity, and baseline search propensity, so a natural question is whether the disparities survive geographic and temporal controls. Table 4 reports logistic regressions on the full dataset. The search disparity is essentially untouched: the adjusted odds ratio on being Black is **2.019** (95% CI 2.000–2.038, $p < 10^{-300}$; $n = 8{,}453{,}751$), almost identical to the raw disparity ratio of 2.10. In other words, the more-than-doubled search rate for Black drivers is *not* an artifact of where or when they are stopped; it holds within counties and within years.

The hit-rate gap behaves very differently. Once county and year fixed effects are included, the adjusted odds ratio on being Black for recovering contraband is **0.986** (95% CI 0.966–1.007, $p = 0.19$; $n = 206{,}866$)—statistically indistinguishable from one. The raw hit-rate deficit for Black drivers is therefore largely a *compositional* phenomenon: Black drivers are concentrated in counties whose overall hit rates are lower, and once that geography is absorbed the within-county hit rates for Black and White drivers are close. This is a concrete instance of the infra-marginality and aggregation concerns emphasized by Anwar and Fang [2] and Simoiu et al. [21]: a raw hit-rate

Table 4: Controlled logistic regressions on the full 2010–2015 dataset. Reported is the odds ratio on `is_black` (Black = 1, White = 0) with year and county fixed effects. The search model is fit over all stops; the hit model over searched drivers only (rare counties collapsed to "Other").

Outcome (model)	OR (is_black)	95% CI	p-value	n
Search $\sim$ is_black $+C$(year) $+C$(county)	2.019	[2.000, 2.038]	$< 10^{-300}$	8,453,751
Contraband $\sim$ is_black $+C$(year) $+C$(county)	0.986	[0.966, 1.007]	0.187	206,866

comparison can mislead when the searched population is unevenly distributed across heterogeneous units.

5 Discussion

5.1 What the two tests jointly imply

The stop-level VOD test and the post-stop outcome test point in complementary directions. At the stop decision, the visibility-mediated contrast the VOD test isolates is small and, once clock time is controlled, runs slightly *against* the classic profiling prediction (OR = 1.045 > 1). We read this cautiously: it is not evidence of discrimination through the specific channel the test probes, but neither is it evidence of its absence in a broader sense, because the VOD design measures only the day/night contrast and is vulnerable to race-correlated differences in who drives after dark [11]. At the search decision, by contrast, the evidence of differential treatment is strong and robust. Black drivers are searched at more than twice the White rate, a disparity that survives county and year controls almost unchanged (adjusted OR = 2.019). The accompanying *lower* hit rate for Black drivers is the classic signature of a lower search threshold, though its statistical force weakens to non-significance under geographic controls.

The contrast between the two decisions is itself informative. The near-invariance of the search disparity to controls, set against the collapse of the hit-rate gap once counties are absorbed, suggests that the strongest signal of unequal treatment in these data lives in the *propensity to search* rather than in the marginal productivity of searches. This is consistent with a large literature finding that post-stop discretion, more than the stop itself, is where racial disparities concentrate [3, 18, 24].

5.2 The role of county and year controls

The differential response of the two outcomes to controls deserves emphasis because it is easy to misread. That the search-rate disparity is stable under fixed effects strengthens the causal interpretation: it is not that Black drivers happen to be stopped in high-search counties: within any given county and year they are still searched about twice as often. That the hit-rate disparity largely disappears under the same controls weakens the outcome-test inference: the pooled hit-rate deficit is substantially a between-county compositional effect rather than a within-county threshold effect. Reporting both the pooled and the controlled estimates is therefore essential; presenting either alone would overstate or understate the case. This sensitivity is exactly why modern treatments favor threshold tests, which model the within-group risk distribution explicitly, over raw hit-rate comparisons [18, 21].

5.3 Limitations

Several limitations bound the interpretation. First, the darkness indicator uses a *single statewide reference* (Raleigh). North Carolina spans roughly 33 minutes of solar time from its eastern to western edge, so stops far from Raleigh are classified with measurement error; this attenuates the darkness coefficient toward one and could partly explain the small adjusted effect. Second, and more seriously, *49.2% of stops lack a recorded time* and are necessarily excluded from the VOD sample. If the propensity to record a time is correlated with driver race, agency, or location, the inter-twilight sample may be selected in ways that bias the darkness estimate; the direction of any such bias is unknown. Reassuringly, the outcome-test patterns in the inter-twilight subsample track those in the full dataset, which argues against gross unrepresentativeness on the dimensions we can check. Third, the VOD test identifies only a visibility-mediated contrast and cannot recover the total level of stop disparity or its mechanism; the endogenous-driving critique means an odds ratio above one need not imply reverse discrimination [11]. Fourth, the outcome test rests on the assumption of homogeneous offending distributions; the infra-marginality problem means unequal hit rates are suggestive but not dispositive proof of bias, and the collapse of the hit-rate gap under controls illustrates precisely this fragility [2, 17, 21]. Finally, `contraband_found` is recorded only for conducted searches, so hit rates carry no information about the unsearched population and are selected by design.

5.4 Directions for stronger identification

Three extensions would sharpen the inference. Assigning each stop a *local* sunset from its county or agency location would reduce the measurement error in the darkness indicator and could be implemented with the OPP county field. Modeling the search decision with a Bayesian *threshold test* [21], which jointly estimates race-specific risk distributions and thresholds, would place the outcome-test intuition on firmer footing than raw hit-rate comparisons and is the current state of the art [18]. And an internal-benchmarking or doubly-robust design [19] would provide a third, methodologically distinct estimate of stop-level disparity to triangulate against the VOD result. We release the full pipeline to make such extensions straightforward.

6 Conclusion

Applying the veil-of-darkness test to 9.49 million North Carolina traffic stops from 2010–2015, we find that once clock time and season are controlled the odds a stopped driver is Black are about 4.5% higher in darkness than daylight (adjusted OR = 1.045, 95% CI 1.024–1.067)—small in magnitude and opposite in sign to the classic profiling prediction, and best read as the absence of the specific day/night signature the test detects. The post-stop outcome test tells a sharper story: Black drivers are searched at 2.10 times the White rate, a disparity that is robust to county and year controls (adjusted OR = 2.019), while their lower contraband hit rate—the textbook signature of a lower search threshold—is real in the pooled data but largely a between-county compositional artifact that fades under geographic controls. The strongest and most control-robust evidence of unequal treatment in these data lies in the decision to search rather than in the visibility contrast at the stop or in the marginal productivity of searches. We state our inter-twilight window explicitly (17:00:27–21:05:08 local clock time), document Raleigh-referenced `astral` sunset/dusk times as the source, and release the code, analysis dataset, and model-output CSVs for full reproducibility.

Data and Code Availability

The source data are the Stanford Open Policing Project standardized North Carolina statewide file (release `2020_04_01`, dataset `druid:yg821jf8611`) [22]. The analysis pipeline (four Python scripts), the inter-twilight analytical dataset (`nc_stops_inter_twilight.csv.gz`, 824,776 rows), the daily Raleigh solar table (`raleigh_sun_2010_2015.csv`), the veil-of-darkness coefficient table (`vod_regression_results.csv`), and the outcome-test tables (`post_stop_rates.csv`, `post_stop_tests.csv`, `post_stop_regressions.csv`) accompany this report. Solar computations used `astral` v3.2 [13]; models were fit with `statsmodels` [20].

Acknowledgments

This work relies on the open data infrastructure of the Stanford Open Policing Project and the open-source scientific Python ecosystem [10, 12, 15, 23].

References

[1] Kate Antonovics and Brian G. Knight. A new look at racial profiling: Evidence from the boston police department. *The Review of Economics and Statistics*, 91(1):163–177, 2009. doi: 10.1162/rest.91.1.163.

[2] Shamena Anwar and Hanming Fang. An alternative test of racial prejudice in motor vehicle searches: Theory and evidence. *American Economic Review*, 96(1):127–151, 2006. doi: 10.1257/000282806776157579.

[3] Frank R. Baumgartner, Derek A. Epp, and Kelsey Shoub. *Suspect Citizens: What 20 Million Traffic Stops Tell Us about Policing and Race*. Cambridge University Press, Cambridge, UK, 2018. doi: 10.1017/9781108553599.

[4] Gary S. Becker. *The Economics of Discrimination*. University of Chicago Press, Chicago, IL, 1957.

[5] Robin S. Engel and Jennifer M. Calnon. Comparing benchmark methodologies for police-citizen contacts: Traffic stop data collection for the pennsylvania state police. *Police Quarterly*, 7(1):97–125, 2004. doi: 10.1177/1098611103257686.

[6] Roland G. Fryer. An empirical analysis of racial differences in police use of force. *Journal of Political Economy*, 127(3):1210–1261, 2019. doi: 10.1086/701423.

[7] Andrew Gelman, Jeffrey Fagan, and Alex Kiss. An analysis of the new york city police department's "stop-and-frisk" policy in the context of claims of racial bias. *Journal of the American Statistical Association*, 102(479):813–823, 2007. doi: 10.1198/016214506000001040.

[8] Sharad Goel, Justin M. Rao, and Ravi Shroff. Precinct or prejudice? understanding racial disparities in new york city's stop-and-frisk policy. *The Annals of Applied Statistics*, 10(1):365–394, 2016. doi: 10.1214/15-AOAS897.

[9] 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.

[10] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with numpy. *Nature*, 585(7825):357–362, 2020. doi: 10.1038/s41586-020-2649-2.

[11] William C. Horrace and Shawn M. Rohlin. How dark is dark? bright lights, big city, racial profiling. *The Review of Economics and Statistics*, 98(2):226–232, 2016. doi: 10.1162/REST_a_00543.

[12] John D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. doi: 10.1109/MCSE.2007.55.

[13] Simon Kennedy. astral: Calculations for the position of the sun and moon. https://astral.readthedocs.io, 2023. Python package, version 3.2.

[14] John Knowles, Nicola Persico, and Petra Todd. Racial bias in motor vehicle searches: Theory and evidence. *Journal of Political Economy*, 109(1):203–229, 2001. doi: 10.1086/318603.

[15] Wes McKinney. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference (SciPy 2010)*, pages 56–61, 2010. doi: 10.25080/Majora-92bf1922-00a.

[16] Nicola Persico. Racial profiling, fairness, and effectiveness of policing. *American Economic Review*, 92(5):1472–1497, 2002. doi: 10.1257/000282802762024593.

[17] Nicola Persico. Racial profiling? detecting bias using statistical evidence. *Annual Review of Economics*, 1(1):229–254, 2009. doi: 10.1146/annurev.economics.050708.143307.

[18] Emma Pierson, Camelia Simoiu, Jan Overgoor, Sam Corbett-Davies, Daniel Jenson, Amy Shoemaker, Vignesh Ramachandran, Phoebe Barghouty, Cheryl Phillips, Ravi Shroff, and Sharad Goel. A large-scale analysis of racial disparities in police stops across the united states. *Nature Human Behaviour*, 4(7):736–745, 2020. doi: 10.1038/s41562-020-0858-1.

[19] Greg Ridgeway and John M. MacDonald. Doubly robust internal benchmarking and false discovery rates for detecting racial bias in police stops. *Journal of the American Statistical Association*, 104(486):661–668, 2009. doi: 10.1198/jasa.2009.0034.

[20] Skipper Seabold and Josef Perktold. Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference (SciPy 2010)*, pages 92–96, 2010. doi: 10.25080/Majora-92bf1922-011.

[21] Camelia Simoiu, Sam Corbett-Davies, and Sharad Goel. The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics*, 11(3):1193–1216, 2017. doi: 10.1214/17-AOAS1058.

[22] Stanford Open Policing Project. Standardized traffic stop data: North carolina statewide (release 2020-04-01). https://openpolicing.stanford.edu, 2020. Dataset druid:yg821jf8611.

[23] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C. J. Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, and Paul van Mulbregt. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272, 2020. doi: 10.1038/s41592-019-0686-2.

[24] Patricia Warren, Donald Tomaskovic-Devey, William Smith, Matthew Zingraff, and Marcinda Mason. Driving while black: Bias processes and racial disparity in police stops. *Criminology*, 44(3):709–738, 2006. doi: 10.1111/j.1745-9125.2006.00061.x.