The Cross-Context Threshold Test: Detecting Discrimination Under Environmental Shifts

Jun Yuan
Digital Technology for Democracy Lab
University of Virginia
uzz3jj@virginia.edu

Xinyue Ye
Department of Geography and the Environment
University of Alabama
xye10@ua.edu

Abstract

We study how threshold tests for detecting discrimination under environmental shifts, focusing on the Veil-of-Darkness (VoD) setting where visibility changes between daylight and darkness. We show that standard threshold tests, when applied separately to daylight and darkness data, violate key assumptions: risk distributions drift across contexts and thresholds fluctuate arbitrarily. We propose a cross-context threshold test that enforces distributional invariance and monotonic threshold decay. Using New York City stop-and-frisk data and synthetic experiments, we demonstrate that this model yields more reliable thresholds, improves bias detection, and aligns with the counterfactual logic of the VoD test. Our framework generalizes to fairness auditing whenever environmental context influences decisions.

1 INTRODUCTION

Auditors tasked with detecting bias in decision-making systems often face two challenges: (1) adapting existing statistical tests to new contexts, and (2) developing tools that make audit outcomes interpretable and actionable. Both remain under-explored. In policing, for example, officers stop cars or pedestrians to search for contraband. The underlying decision process can be viewed as a binary classifier: each individual has a latent probability of carrying contraband, and an officer chooses a threshold above which to initiate a stop. Observed stop outcomes (successful versus unsuccessful stops) provide indirect evidence of these thresholds. A recent work introduced the *threshold test*, which estimates decision thresholds for racial groups from observed stop counts, hit counts, and census-based population proportions (Pierson et al., 2018) . By jointly estimating both group-level risk distributions and thresholds, the threshold test addresses the well-known problem of inframarginality (Arnold et al., 2018; Ayres, 2002; Engel, 2008; Pierson et al., 2020). A complementary strategy is the *veil-of-darkness* (VoD) test (Grogger and Ridgeway, 2006; Worden et al., 2012; Knode et al., 2024), which exploits a natural experiment: after sunset, reduced visibility makes it more difficult for officers to identify a person's race before deciding to stop them. The VoD test assumes that underlying risk distributions remain stable across twilight, so that any observed changes in stop rates or thresholds can be attributed to racial bias. Recent work often combines threshold and VoD tests to study racial disparities in policing (Pierson et al., 2020; Knode et al., 2024). Some researchers extend this by applying threshold tests separately to daylight and darkness subsets, interpreting differences in estimated thresholds as evidence aligned with VoD logic—that racial disparities diminish when race is visually obscured (Pierson et al., 2020). However, this direct application overlooks key VoD assumptions: (1) the risk distribution for any racial group should not shift abruptly with visibility, and (2) thresholds should converge under reduced visibility. When these assumptions are violated, the test results are unstable or misleading.

To address this gap, we propose the *cross-context threshold test* (CCTT), which adapts the threshold test to explicitly account for VoD conditions. Our model enforces both distributional invariance across daylight and darkness and monotonic threshold decay under reduced visibility. Through controlled simulation studies, we demonstrate that CCTT more accurately recovers ground-truth thresholds and achieves

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

higher recall in detecting bias than the standard threshold test. Applied to New York City stop-and-frisk data, CCTT uncovers additional racial disparities and better aligns with the counterfactual logic of the VoD framework. Beyond policing, this approach generalizes to other high-stakes domains where context shifts alter decision thresholds but not underlying risk distributions.

2 BACKGROUND

2.1 Threshold Tests for Discrimination

Outcome tests compare success rates (e.g., hit rates) across groups but suffer from the inframarginality problem: differences in hit rates do not necessarily imply differences in decision thresholds (Ayres, 2002; Engel, 2008; Simoiu et al., 2017). Threshold tests address this limitation by jointly estimating group-level risk distributions and the thresholds applied to those distributions (Pierson et al., 2018). The model infers how officers set stop thresholds across racial groups using observed stop counts, hit counts, and population proportions. This approach has become a leading method for identifying disparate treatment in policing data and has been applied at scale across jurisdictions in the United States (Pierson et al., 2020).

2.2 The Veil-of-Darkness (VoD) Framework

The veil-of-darkness test, introduced by Grogger and Ridgeway (2006), provides a complementary quasi-experimental approach. It exploits the natural transition around sunset: after dark, reduced visibility makes it more difficult for officers to discern a pedestrian's or driver's race before deciding to stop them. The core assumption is that the underlying risk distribution of contraband possession does not change abruptly around twilight. Thus, any observed changes in stop rates or thresholds across daylight and darkness can be interpreted as evidence of bias in policing practices (Knode et al., 2024; Worden et al., 2012).

2.3 The Gap

Recent work (Pierson et al., 2020) has attempted to combine these approaches by applying threshold tests separately to daylight and darkness stops, interpreting threshold differences as evidence aligned with VoD logic. However, this direct application violates key assumptions. Risk distributions are estimated independently across day and night, leading to instability, and thresholds may fluctuate arbitrarily. As we show in Section 4, this motivates the *cross-context threshold test* (CCTT), which explicitly enforces distributional invariance and monotonic threshold decay across visibility conditions.

3 THE CROSS-CONTEXT THRESHOLD TEST (CCTT)

To analyze how environmental context influences bias detection, we formalize three components of our framework: (1) data curation under contextual conditions, (2) the classification task being audited, and (3) the bias testing method. This structure allows us to clearly identify where context enters the analysis and how it shapes the interpretation of bias metrics.

3.1 Problem Setup

Let $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{N}$ be a dataset of N decision instances, where $\mathbf{x}_i$ is the feature vector and y_i is the observed decision outcome (e.g., whether a stop yielded contraband). In many applications, we analyze a curated subset of $\mathcal{D}$ that meets specific contextual conditions. In our case, the veil-of-darkness (VoD) setting. We denote this curated dataset as:

$$\mathcal{D}_K = f(\mathcal{D}, K)$$

where K encodes two kinds of auxiliary conditions:

1. **Bias-analysis scope** (K_{bias}): the subset of instances and outcomes under analysis (e.g., a specific region, time window, or racial groups of interest).
2. **Distribution priors** (K_{dist}): assumptions about underlying population composition or distributions.

If no auxiliary conditions are applied, $\mathcal{D}_K = \mathcal{D}$.

The decision process being audited is represented as a function h mapping features to outcomes:

$$h = \mathcal{M}(\mathcal{D}_K)$$

Here, h may be as simple as a lookup from the dataset (when outcomes are directly observed) or a model that generalizes to new instances from a training paradigm $\mathcal{M}$. In this work, we treat h as a lookup function to focus on auditing observed historical decisions.

A **bias test** computes a group-level metric $\mathbf{t} = \{t_{p_1}, \ldots, t_{p_s}\}$ for each of s groups $\mathbf{p} = \{p_i\}_{i=1}^{s}$:

$$\mathbf{t} = \mathcal{B}(h, \mathcal{D}_K)$$

Examples of t_{p_i} include positive rates (PR), true positive rates (TPR), or decision thresholds. A statistically significant difference among these values suggests potential bias. When the bias test is adapted to a specific context K, we write $\mathcal{B}_K$.

3.2 Assumptions

In this section, we describe the assumptions K for the VoD threshold test to define the $\mathcal{B}_K$.

Assumption 1: Risk (signal) distribution for each racial group does not change from daylight to darkness during the intertwilight period.

Assumption 2: Police always find it more difficult to distinguish race at darkness than daylight due to limited visibility.

The Assumption 1 reflects a simple intuition that while group risk distributions may evolve over long time period (e.g., years of demographic or policy change), it is reasonable to treat them as stable across the short temporal window of twilight. This is consistent with the VoD literature (Worden et al., 2012), which recommends restricting analysis to a narrow twilight window where population characteristics are held constant except for visibility. The implication of Assumption 2 may be interpreted as the thresholds are more similar (i.e., smaller variance) at darkness than daylight among racial groups. Direct enforcement of this assumption is difficult. As a practical alternative, we impose that the threshold in darkness is less than or equal to its daylight counterpart. This condition reflects the interpretation that officers exercise a less selective (i.e., more cautious) decision rule at night for the same racial group. Such behavior is consistent with the VoD literature (Grogger and Ridgeway, 2006) which suggests that reduced visibility leads officers to rely on lower stopping thresholds. When thresholds are restricted to the interval $[0, 1]$, this condition also implies Assumption 2 (proof provided in the Supplementary Material). Importantly, lower thresholds do not necessarily translate into a higher number of stops, as stop counts also depend on factors such as street population and patrol exposure.

We first validate the proposed cross-context VoD assumptions under controlled simulation. Then we apply them to more complex real-world policing data and adapt models from prior work.

3.3 Simulation Study

For simulation, we introduce the simplified CCTT. Each group-precinct pair is assigned a single latent risk distribution,

$$Z\ g, p \sim \mathcal{N}(\mu_{g,p}, \sigma^2), \tag{1}$$

Shared across lighting condition (i.e., day and night). Threshold follows

$$t^{\text{day}}_{g,p} \sim \mathcal{N}(\mu_{g,p}, 1) \tag{2}$$

$$t^{\text{night}}_{g,p} = t^{\text{day}}_{g,p} - \delta_{g,p}, \quad \delta_{g,p} \geq 0 \tag{3}$$

This simplified specification, including the use of a Normal distribution for the risk distribution, is intentional to isolate the roles of distributional invariance and monotone decay. To highlight the impact of applying VoD constraints on existing work, later sections adopt a Logit-Normal Mixture model instead of the Normal distribution, as it is more commonly used in prior threshold-based work (Pierson et al., 2018, 2020). Here, we assess identifiability and bias detection performance on simulated data.

For the context $c \in \{\text{day}, \text{night}\}$. Let $N^c_{g,p}$ denotes the number of people in a racial group g and police precinct p at the context c. The stop count is

$$S^c_{g,p} \sim Binomial(N^c_{g,p}, P(Z_{g,p} > t^c_{g,p})) \tag{4}$$

The hit count is

$$H^c_{g,p} \sim Binomial(S^c_{g,p}, E[(Z_{g,p} > t^c_{g,p}]) \tag{5}$$

We report the following priors used for this simulation study. $Z_{g,p} \sim \mathcal{N}(0, 1.5^2)$ $t^{\text{day}}_{g,p} \sim \mathcal{N}(0, 2^2)$ $\delta_{g,p} \sim HalfNormal(0.2, 0.4^2)$

A comparable standard threshold test (TT) does not impose conditions of shared risk distribution or threshold decay. For TT, we define and estimate the risk distributions $Z^c_{g,p}$ and thresholds $t^c_{g,p}$ separately for each context. We omit the cross-context constraint (Equation 3), while keeping the remaining steps (Equations 4 and 5) unchanged.

3.3.1 Generating Ground Truth

We simulate 50 precincts for two racial groups (e.g., White and Black). Risk signals for class $Y \in \{\text{guilty} = 1, \text{innocent} = 0\}$ are generated as

$$X \mid Y = 0 \sim \mathcal{N}(0, 1), \quad X \mid Y = 1 \sim \mathcal{N}(\mu_1, 1),$$

$\mu_1 \in [1.5, 2.5]$. This construction ensures a separation between innocent and guilty individuals while producing realistic hit rates (10%–20%) when thresholds are applied.

Daytime thresholds are assigned at the precinct level, with a race-specific gap:

$$t^{\text{day}}_{g=\text{white},p} \sim U(0.35, 0.55), \tag{6}$$

$$t^{\text{day}}_{g=\text{black},p} = t^{\text{day}}_{g=\text{white},p} \pm \Delta_p, \quad \Delta_p \sim U(0.05, 0.15). \tag{7}$$

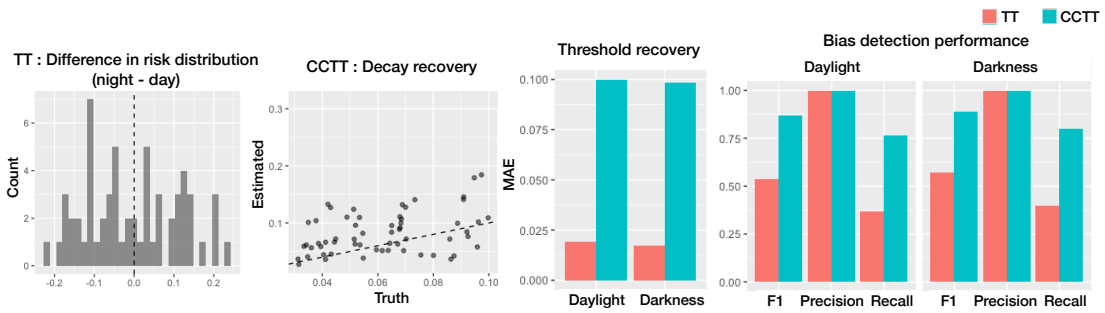

Figure 1: **Simulation study outcome.** From left to right, 1) TT generates different day night risk distribution. 2) CCTT estimates the decay is positively correlated with true decay. 3) CCTT estimated the ground truth thresholds worse than TT. However, TT appears to overfit due to its flexibility in fitting separate distributions. 4) The F1 and recall scores shows that CCTT is better at detecting the true bias than TT.

Nighttime thresholds are defined by applying a non-negative decay to the daytime thresholds:

$$t_{g,p}^{\text{night}} = t_{g,p}^{\text{day}} - \delta_{g,p}, \quad \delta_{g,p} \sim U(0.03, 0.10). \tag{8}$$

The ground-truth population sizes $N_{g,p}^{\text{day}}$ and $N_{g,p}^{\text{night}}$ are drawn independently from a Poisson distribution with mean 2000. For each individual, if $z > t$ the model records a stop S, and if additionally $Y = 1$, a hit H.

This setup guarantees that risk distributions are invariant across day and night while thresholds monotonically decrease in darkness. It also embeds true racial disparities through precinct-level gaps, providing a benchmark for evaluating whether the threshold test (TT) and our cross-context threshold test (CCTT) can correctly recover model parameters and detect bias.

3.3.2 Results

The simulation took approximately 5 minutes on a 32 GB RAM Intel MacBook Pro using R. The results are shown in Fig. 1. Fig. 1.1 shows that the TT estimated the day and night risk distributions differ significantly, measured by the difference of μ between day and night. The differences are arbitrarily positive or negative but the actual difference should be negative. Fig. 1.2 shows that the CCTT's estimation of threshold decay is positively correlated to the true decay. Fig. 1.3 shows that the TT recovers thresholds with a lower Mean Absolute Error (MAE), better than CCTT. But considering Fig. 1.1, this indicates TT is over-fitted due the flexibility in two risk distributions. So relying on MAE is misleading in determining model performance. Fig. 1.3 shows that actual performance in bias detection. TT only detects bias when it's absolutely sure (i.e., high precision), so it misses many true biased precincts. CCTT has larger recall and overall higher F1, indicating better performance as bias detection method.

3.4 Logit-Normal Mixture Model

In this section, we describe modifications to the threshold test models introduced in prior work. The original threshold test represents group-level risk distributions using Beta distributions (Simoiu et al., 2017). Because the Beta distribution is naturally supported on the interval $[0, 1]$, this formulation also constrains estimated thresholds to lie within that range. Bayesian estimation for beta distributions is time-consuming. A fast threshold test is proposed to utilize a logit-normal mixture model as a proxy for the Beta distribution, thereby speeding up the test by approximately 40 times while achieving the same test results (Pierson et al., 2018). They apply the threshold test to the NYC stop-and-frisk dataset. Each instance in the dataset represents a pedestrian stop record at a certain time and location in New York City. Police search (frisk) decision is recorded (searched or not searched), as well as the outcome (a hit or not a hit). A hit refers to the fact that the police are correct in their search decisions (i.e., the discovery of contraband). They also propose a threshold test for search (frisk) decision and a threshold test for stop decision by estimating the racial proportion of pedestrians in each police precinct using Census data.

We focus on the threshold test for *stop decisions*, as opposed to searches, because they are more directly influenced by visibility: an officer may not discern a pedestrian's race at a distance in darkness, but will almost always do so before deciding to frisk.

Here, we introduce the complete fast threshold test. Y

is the result of stop (guilty = 1, innocent = 0) based on the risk probability z.

$$Y = Bernoulli(z) \tag{9}$$

X is the risk signal of the person (pedestrian) that follows a normal distribution given that they are guilty or innocent.

$$X|Y = 0 \sim N(\mu_0, \sigma_0) \tag{10}$$

$$X|Y = 1 \sim N(\mu_1, \sigma_1) \tag{11}$$

$$Z = g(x; \mu_0, \sigma_0, \mu_1, \sigma_1) = P(Y = 1|X = x) \tag{12}$$

$g(x)$ is the risk function that maps a real value x to the probability of the risk bounded between 0 and 1, denoted as Z. Determining the risk distribution for each race is achieved by estimating μ_0, σ_0, μ_1, σ_1 per race. We can reduce the number of parameters to be estimated following prior work (Pierson et al., 2018).To ensure heteroskedasticity for g, we constrain $\sigma_0 = \sigma_1$. This homoskedasticity condition ensures that a higher signal X corresponds to a higher risk probability Z (i.e., monotonicity); otherwise, the likelihood may become non-monotonic and thresholds may be non-identifiable. Without loss of generality, we set $\mu_0 = 0$ and estimate only μ_1. Setting $\mu_0 = 0$ for each group is a standard location normalization that does not affect cross-group comparability. Only the difference between μ_0 and μ_1 determines the mapping from X to risk Z, and thresholds are estimated in Z-space, which is invariant to affine shifts. While estimating the distribution, we also need to estimate the stop threshold t per race per police region simultaneously. We fit the model to the ground truth data, such as hit count and stop count, to determine the best model parameter values.

Step 1: For each race, we fit the stop count and hit count per police region:

Let S be the stop count.

$$S \sim Binomial(\text{population}, \text{stop rate}) \tag{13}$$

Stop rate can be expressed as the probability of risk above the stop threshold t:

$$\begin{aligned} P(Z > t) =& P(X > g^{-1}(t)|Y = 1)P(Y = 1) + \\ & P(X > g^{-1}(t)|Y = 0)P(Y = 0) \end{aligned} \tag{14}$$

Let H be the hit count.

$$H \sim Binomial(\text{stop count}, \text{hit rate}) \tag{15}$$

Hit rate can be expressed as the expectation of the risk above the stop threshold t:

$$E[Z|Z > t] = \frac{P(X > g^{-1}(t)|Y = 1)P(Y = 1)}{P(Z > t)} \tag{16}$$

Step 2: Across races, we fit the stop count per police region:

Let $\mathbf{N}$ be a vector that contains the count of stops for races. Let N to be the sum of $\mathbf{N}$, the total count of stops between races in one precinct. Let θ be a vector that contains the stop probabilities for races.

$$\mathbf{N} \sim Multinomial(N, \theta) \tag{17}$$

The stop probability can be estimated from the stop rate and the US Census racial proportion, denoted as c.

For example, the stop probability for the white race, θ_w, follows:

$$\theta_w \propto P(Z_w > t_w)c_w \tag{18}$$

c_w is the racial proportion of whites. The white stop threshold t_w and white stop rate $P(Z_w > t_w)$ are estimated in Step 1.

Incorporating the assumptions in the fast threshold test model is simple. For Assumption 1, We adapt the model in following steps: We repeat *step* 1 and 2 for the stop and hit counts for daylight and darkness data. The stop thresholds for daylight stop and hit data and darkness stop and hit data are estimated in each run. But they share the parameters in function g in step 1 so the risk distribution remains one per race regardless of daylight and darkness.

For Assumption 2, we use an additional constraint to ensure threshold at darkness is always smaller than the threshold at daylight. First, the daylight threshold is sampled, same as before, from a standard normal distribution.

$$t_{daylight} \leftarrow normal(0, 1) \tag{19}$$

Then, a parameter t_{decay} is sampled as the decay from the daylight to darkness, also from a standard normal distribution. We set the lower bound for t_{decay} as 0.

$$t_{decay} \leftarrow normal(0, 1),\ t_{decay} \geq 0 \tag{20}$$

Then we set the threshold at darkness at the same race and location (e.g., precinct),

$$t_{darkenss} = t_{daylight} - t_{decay} \tag{21}$$

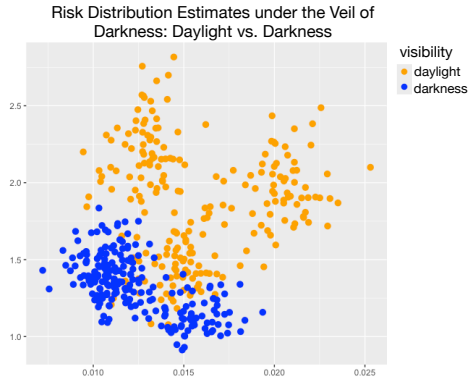

Figure 2: **Issue of threshold test.** Estimated risk distribution parameters projected in 2D scatter plot. The vast difference between the orange and blue scatter plots indicates that the risk distributions between daylight and darkness are estimated to be very different, which does not align with the VoD assumption for distribution invariance per racial group.

All thresholds are then transformed to stay between 0 and 1 using a normal linkage function.

4 APPLICATION: NYC STOP-AND-FRISK DATA

We apply CCTT (described in Section 3.4) on New York City stop-and-frisk records from 2008–2012. The dataset (Pierson et al., 2018) contains 723,553 stops flagged as suspected contraband possession, along with 2010 Census estimates of racial composition by precinct. The distribution priors used in these threshold tests estimate separate risk distributions for each race within each precinct, capturing precinct-level differences in policing practices. Using the same dataset and distribution priors as the basis of our experiments allows us to validate limitations, assess consistency, and directly compare our results with prior work. Notably, the official NYC stop-and-frisk portal provides data spans from 2003 to recent years. In more recent records, the rank of the police officer is also included.

Data and Curation We further create the curated data $\mathcal{D}_K$ for VoD. Since VoD is an effect that police stopping bias is reduced under low visibility (i.e., darkness). This implies that we want to include a stop event that may occur in either darkness or daylight in a given analysis scope (e.g., time-span and geo-boundary). In other words, an event's counterfactual event *has a chance to occur*. Such a counterfactual event only differs from the original event in visibility (i.e., darkness or daylight). For a time span of February 2018, and the geo-boundary of NYC, a stop at 8 PM local time would be impossible to occur in the daylight, unfit for understanding the veil of darkness on police stopping. However, if the time span changes to the year 2018, such a stop can be included in the VoD analysis since 8 PM in May can be at daylight. Such nuanced data curation based on analysis scope is crucial to the rigorousness of downstream statistical testing. To reflect the VoD data condition, We provide the curated VoD data (146,491 records) [1] for promoting reproducibility and transparency. The VoD data $\mathcal{D}_K$ is about 20 percent of the total $\mathcal{D}$. We follow the tutorial of a recent work (Knode et al., 2024) for the data curation process. Here are the four steps: 1) We split the data by year. 2) For each year, we identify the stops during intertwilight. 3) For the stop in the intertwilight, we define stops between sunset and dusk as ambiguous stops, stops before sunset as daylight stops, and after dusk as darkness stops. 4) We exclude the ambiguous stops and aggregate the remaining stops across the year from 2008 to 2012. The data process took about 5 hours on a 32 GB RAM, Intel MacBook Pro. Note that one could choose not to split the data by year but set the time-span from the first day of 2008 to the last day of 2012, which will likely increase the record count of VoD data. As the possible range of the sunset and dawn time is wider from a longer time span than any sub-time span. However, larger time-span increases the computational cost.

Violation of Standard Threshold Test. We apply the threshold test on both daylight and darkness data. The estimated group-level risk distributions differ sharply rather than overlaying between visibility conditions (Fig. 2), violating the VoD assumption that visibility blurs group signals without altering underlying risk distributions.

4.1 Visual Audit Dashboard

As practical auditing may involve numerous contexts, traditional static plots is insufficient for comparing the effect of assumptions and decisions. We demonstrate the application of our cross-context threshold test framework in a flexible and extendable visual dashboard (Fig. 3) for New York City stop-and-frisk veil-of-darkness (VoD) analysis. The dashboard was developed to address the oversight challenge that human auditors must often compare results from multiple bias detection methods, interpret the magnitude and direction of disparities, and decide which jurisdictions warrant further investigation. Prior work (Pierson et al., 2018) uses a static threshold plot to show the pattern of thresholds between two dif-

[1] Code to reproduce the experiments is available online. `https://github.com/junyuan-ai/cctt-threshold-test`

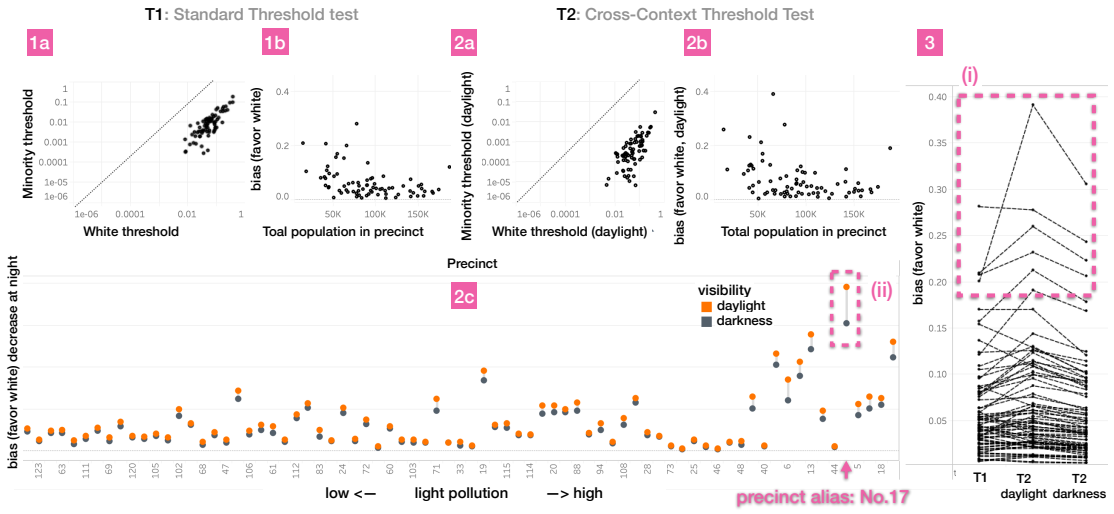

Figure 3: **Interactive audit dashboard for veil-of-darkness analysis.** Panels show: threshold plots (1a,2a), bias plots (1b, 2b), contrastive bias plot (2c), and consensus plot (3), comparing bias judgment based on thresholds derived from the original TT and the CCTT. Plots correspond to: 1a, 1b: threshold test without context. 2a, 2b, 2c: threshold test under VoD context 3: summarizes precinct-level bias values from all three tests using parallel coordinates, enabling auditors to see which precinct jurisdictions are flagged under each method.

ferent groups (e.g., white and minority racial groups). Threshold plots (Fig.3(1a, 2a)) directly encode the thresholds of two groups for the same police precinct, but the degree of bias can only be visually perceived as the distance from the dot to the diagonal line (i.e., dots falling on the diagonal line indicates the equal thresholds the precincts use for stop decision). The static threshold plot also does not support flexible exploration and iterative investigation of the context or the instances (e.g., police precincts) under investigation, which highlight the importance of interactivity in an audit dashboard. Note that the Fig.3(1a) reproduces the Figure 4 (left) in prior work (Pierson et al., 2018), omitting the size difference of dots (which encoded the population size). Fig.3(2a) is the cross-context threshold plot with the exponential axes scaling same as Fig.3(1a). Comparing the figures (Fig.3(1a, 2a)) tells us that the threshold pattern and bias conclusion remains the same between the original threshold test and CCTT–the police favors the white group.

We introduce three additional plots (bias plot, contrastive bias plot, and consensus plot) to support direct encoding of bias and cross-context comparison. The bias plot (Fig.3(1b, 2b)) directly encodes the magnitude of disparity as the vertical distance from a fairness baseline. The contrastive bias plot (Fig.3(2c)) focuses on the change in disparity between daylight and darkness (i.e., the threshold decay). The consensus plot (Fig.3(3)) summarizes results from multiple bias test outcomes, allowing auditors to identify precincts flagged as biased under at least one case.

The dashboard supports interactive filtering and cross-highlighting across plots, enabling an auditor to, for example, select precincts with bias above a chosen threshold in one plot and immediately see their status in others. This allows auditors to discover precincts where disparities are both large and consistent across bias test methods (Fig.3(i)).

For the contrastive bias plot (Fig.3(2c)), we bring in light pollution data to infer the effect of the veil-of-darkness (VoD) from underlying policing practices. We use the average masked composite from the VIIRS Nighttime Lights (VNL V2) dataset (Elvidge et al., 2021), which provides a stable measure of persistent anthropogenic light. Transient sources such as fires and boats are removed through masking, ensuring data reflect long-term urban light pollution. This dataset balances between spatial resolution (∼500 m) and file size (∼262 MB), and is sufficient for precinct-level aggregation in New York City.

In Fig. 3(2c), the precinct at the far right corresponds to Times Square, which exhibits the highest light pollution. According to the VoD logic, policing outcomes should be fairer at night, since reduced visibility implies that officers rely on lower thresholds across racial groups. However, this does not mean that they are affected equivalently due to light pollution. We for-

malize stop bias within an explanatory causal model (ECM) [2], which specifies bias as the causal effect of visibility and policing practice. The visibility affects the police practice, then police practice affects the stop bias, so changes in stop bias cannot be interpreted in isolation. In structural form,

$$\text{Visibility} \longrightarrow \text{Police Practice} \longrightarrow \text{Stop Bias}$$

In precincts with high levels of light pollution, the difference between daylight and darkness is minimal, yet we observe substantial reductions in stop bias (e.g., precinct alias No.17). Since visibility effects are unlikely to explain such drops, these reductions are more plausibly associated with changes in policing practices. This does not imply that policing in these contexts is unbiased or fair. A reduction in disparity under specific conditions should be interpreted only as evidence of practice shifts, not as the absence of bias (e.g., precinct No.17 has large bias even after bias decrease at night). In precincts with low levels of light pollution, where the transition from daylight to darkness substantially reduces visibility (e.g., precinct on the far left of the Figure 3(2c)), one would expect bias to diminish if officers relied less on racial cues. Yet our results show little or no reduction. This persistence suggests that policing practices may actively counteract the expected visibility effect, sustaining disparities even when race is less identifiable. Using ECM framing, we can disentangle environmental influences from institutional practices and prevent over-interpretation of statistical outcomes as straightforward evidence of fairness. Our study does not claim full causal identification, but to align the cross-context threshold test framework with broader causal inference literature and clarify how observed disparities should be interpreted.

5 DISCUSSION

While our study emphasized visibility as the primary environmental context shaping bias detection, the framework naturally extends to other contextual dimensions within policing and to domains beyond law enforcement. Within policing, institutional roles create additional layers of decision-making. Patrol officers, supervisors, and specialized units differ in authority, oversight, and exposure. Similar to the veil-of-darkness (VoD) setting, the risk distribution of racial groups should remain invariant across these organizational contexts. Differences in estimated thresholds therefore reflect institutional practices rather than population risk heterogeneity. Moreover, environmental and organizational factors can intersect: for example, supervisors may apply different daylight versus darkness thresholds than frontline officers. Such intersections enrich contextual auditing.

The CCTT framework extends beyond policing. In AI model auditing, visibility can be reinterpreted as the amount of context available in prompts: high-context prompts expose more features, while low-context prompts obscure them, analogous to visibility changes in policing. CCTT allows auditors to separate shifts in underlying "risk distributions" from shifts in decision thresholds, clarifying whether disparities stem from model mechanisms or input conditions. In medical imaging, differing image resolutions create challenges: low-quality scans may obscure signals from similar disease (e.g., pneumonia and lung cancer), shifting diagnostic thresholds while disease prevalence remains constant. By curating datasets across imaging conditions and holding latent distributions fixed, threshold analysis can reveal systematic decision shifts attributable to degraded visibility. These examples show that the CCTT framework is flexible and generalizable. Furthermore, embedding context directly into threshold tests yields more interpretable and actionable evidence.

CCTT has several limitations. First, threshold tests require specifying a parametric form for the risk distribution. Choices such as Normal, Beta, or logit-normal mixtures may still fail to capture the true underlying distributions. Second, while CCTT runs in comparable time to the fast threshold test, it still requires minutes of MCMC sampling in R. This constrains flexibility and reduces the ability to redefine contexts interactively in an auditing dashboard.

6 CONCLUSION

We introduce the Cross-Context Threshold Test (CCTT), a framework for adapting threshold tests to settings where environmental or institutional context influences decision-making. Standard threshold tests, when applied separately to daylight and darkness, violate the assumptions of the veil-of-darkness (VoD) framework by allowing risk distributions to drift arbitrarily and thresholds to fluctuate inconsistently. CCTT addresses these issues by explicitly enforcing distributional invariance across contexts and monotonic threshold decay, aligning statistical tests with the counterfactual logic of VoD. Through simulation studies, we show that CCTT yields higher recall and F1 scores while maintaining perfect precision in detecting biased precincts compared to the standard threshold test. Applied to New York City stop-and-frisk data,

[2]While we adopt ECM as a conceptual tool, we do not claim that our observational design fully identifies all causal effects (see Banitz et al. (2022) for caveats of model-derived explanation).

CCTT uncovered additional instances of racial disparities that are overlooked by existing approaches. We also provide a visualization tool to illustrate threshold shifts and cross-context comparisons, highlighting how these methodological advances can be interpreted for oversight. While this paper focused on visibility as an environmental context, the framework extends naturally to organizational settings (e.g., officer rank or unit) and to other domains where context alters decision thresholds without changing underlying risk distributions. Examples include medical imaging under varying resolutions and AI auditing under different prompt conditions. By embedding contextual assumptions directly into threshold tests, CCTT advances both the methodological rigor and the practical utility of fairness audits.

Acknowledgements

We thank the anonymous reviewers for their valuable feedback. We also thank members of the Digital Technology for Democracy Lab at the University of Virginia and Xinyue Ye's research group for helpful discussions and suggestions. This work was supported in part by the University of Virginia.

References

David Arnold, Will Dobbie, and Crystal S Yang. Racial bias in bail decisions. *The Quarterly Journal of Economics*, 133(4):1885–1932, 2018.

Ian Ayres. Outcome tests of racial disparities in police practices. *Justice research and Policy*, 4(1-2):131–142, 2002.

Thomas Banitz, Maja Schlüter, Emilie Lindkvist, Sonja Radosavljevic, Lars-Göran Johansson, Petri Ylikoski, Rodrigo Martínez-Peña, and Volker Grimm. Model-derived causal explanations are inherently constrained by hidden assumptions and context: The example of baltic cod dynamics. *Environmental Modelling & Software*, 156:105489, 2022.

Christopher D Elvidge, Mikhail Zhizhin, Tilottama Ghosh, Feng-Chi Hsu, and Jay Taneja. Annual time series of global viirs nighttime lights derived from monthly averages: 2012 to 2019. *Remote Sensing*, 13(5):922, 2021.

Robin S Engel. A critique of the "outcome test" in racial profiling research. *Justice Quarterly*, 25(1): 1–36, 2008.

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.

Jedidiah L Knode, Scott E Wolfe, and Travis M Carter. Pulling back the veil of darkness: A proposed road map to disentangle racial disparities in traffic stops, a research note. *Criminology*, 62(2): 364–375, 2024.

Emma Pierson, Sam Corbett-Davies, and Sharad Goel. Fast threshold tests for detecting discrimination. In *International conference on artificial intelligence and statistics*, pages 96–105. PMLR, 2018.

Emma Pierson, Camelia Simoiu, Jan Overgoor, Sam Corbett-Davies, Daniel Jenson, Amy Shoemaker, Vignesh Ramachandran, Phoebe Barghouty, Cheryl Phillips, Ravi Shroff, et al. A large-scale analysis of racial disparities in police stops across the united states. *Nature human behaviour*, 4(7):736–745, 2020.

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

Robert E Worden, Sarah J McLean, and Andrew P Wheeler. Testing for racial profiling with the veil-of-darkness method. *Police Quarterly*, 15(1):92–111, 2012.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [No]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [Yes]

   (b) The license information of the assets, if applicable. [Yes]

   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes]

   (d) Information about consent from data providers/curators. [Not Applicable]

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

The Cross-Context Threshold Test: Detecting Discrimination Under Environmental Shifts: Supplementary Materials

A Proof in Section 3.2 Assumption 2

We aim to show that thresholds across races exhibit reduced variance at night compared to daytime.

Let the daytime thresholds be

$$t = (t_1, t_2, \ldots, t_n)^\top \in [0,1]^n,$$

and the corresponding nighttime thresholds be

$$t' = (t'_1, t'_2, \ldots, t'_n)^\top,$$

where for each race i,

$$t'_i = a_i t_i, \quad a_i > 0.$$

Here a_i represents the ratio of the nighttime threshold to the daytime threshold for race i.

We assume that nighttime thresholds are not inflated overall relative to daytime thresholds:

$$\|t'\| \leq \|t\|, \quad \text{or equivalently,} \quad 0 \leq \frac{\|t'\|}{\|t\|} \leq 1. \tag{22}$$

To quantify cross-racial disparity, we define the pairwise variance functional

$$V(t) = \sum_{i<j} (t_i - t_j)^2.$$

The corresponding nighttime variance is

$$V(t') = \sum_{i<j} (t'_i - t'_j)^2 = \sum_{i<j} (a_i t_i - a_j t_j)^2.$$

The ratio between night and day variances is then

$$R = \frac{V(t')}{V(t)} = \frac{\sum_{i<j} (a_i t_i - a_j t_j)^2}{\sum_{i<j} (t_i - t_j)^2}. \tag{23}$$

Each term in the numerator of (23) scales the daytime difference $(t_i - t_j)^2$ by a factor that depends smoothly on a_i and a_j. Since all weights are positive and the function is quadratic in the a_i's, the overall ratio R must lie between the smallest and largest squared scaling factors:

$$\boxed{\min_i a_i^2 \;\leq\; R \;\leq\; \max_i a_i^2.} \tag{24}$$

If $0 < a_i < 1$ for all i, then both bounds in (24) are strictly less than one, implying

$$V(t') < V(t).$$

Therefore, the cross-racial threshold variance decreases at night, indicating that nighttime thresholds are more homogeneous across races.