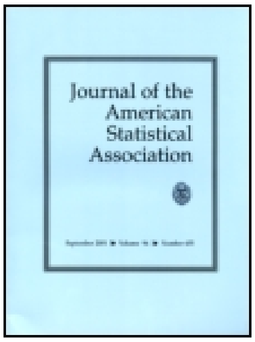

Journal of the American Statistical Association

Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Doubly Robust Internal Benchmarking and False Discovery Rates for Detecting Racial Bias in Police Stops

Greg Ridgeway & John M. MacDonald

To cite this article: Greg Ridgeway & John M. MacDonald (2009) Doubly Robust Internal Benchmarking and False Discovery Rates for Detecting Racial Bias in Police Stops, Journal of the American Statistical Association, 104:486, 661-668, DOI: 10.1198/jasa.2009.0034

To link to this article: https://doi.org/10.1198/jasa.2009.0034

Published online: 01 Jan 2012.

Doubly Robust Internal Benchmarking and False Discovery Rates for Detecting Racial Bias in Police Stops

Greg RIDGEWAY and John M. MACDONALD

Allegations of racially biased policing are a contentious issue in many communities. Processes that flag potential problem officers have become a key component of risk management systems at major police departments. We present a statistical method to flag potential problem officers by blending three methodologies that are the focus of active research efforts: propensity score weighting, doubly robust estimation, and false discovery rates. Compared with other systems currently in use, the proposed method reduces the risk of flagging a substantial number of false positives by more rigorously adjusting for potential confounders and by using the false discovery rate as a measure to flag officers. We apply the methodology to data on 500,000 pedestrian stops in New York City in 2006. Of the nearly 3,000 New York City Police Department officers regularly involved in pedestrian stops, we flag 15 officers who stopped a substantially greater fraction of black and Hispanic suspects than our statistical benchmark predicts.

KEY WORDS: Propensity score weighting; Racial profiling; Risk management systems.

1. INTRODUCTION

Race is at the forefront of most discussions of police behavior (Kennedy 1997; Russell-Brown 1998). Numerous departments face expensive civil litigation as a result of high profile police use of force incidents and allegations of systemic patterns of racially biased police practices. These include United States Department of Justice oversight of settlement agreements in several cities including Los Angeles, Washington DC, and Cincinnati. These agreements often involve extensive data collection efforts and analyses to assess evidence of racially biased policing. Whereas the data collection efforts are producing large, detailed datasets, the extent to which current methodologies can adequately capture bias in police decisions on which citizens to stop and question (e.g., racial profiling) and other routine activities remains a matter of debate (National Research Council 2003). Plaintiffs seek methods that can pinpoint problems in departments and seek remedies. Departments are eager to find civil liability risks, such as problem officers, before any incidents arise.

There have been many efforts to assess whether entire police forces have racially biased practices. The majority of methods used to assess biased police behavior focus on comparisons of a given police department's rate of stops, searches, or arrests of nonwhites against some form of external benchmark (Fagan and Davies 2000; Fridell 2004; Gelman, Fagan, and Kiss 2007). Studies have compared the race distribution of individuals that officers have stopped with the race distribution of residents reported in the census (Steward 2004; Weiss and Grumet-Morris 2005), the race distribution of not-at-fault drivers in traffic accidents (Alpert, Smith, and Dunham 2004), the race distribution of those cited using race blind detection such as photographic stoplight enforcement (Montgomery County Department of Police 2002) or by aerial patrols (McConnell and Scheidegger 2001), and the race distribution of a sample of drivers or pedestrians in public (Lamberth 1994). Studies have also focused on ''outcomes tests'' from police searches conducted during motor vehicle stops (Knowles, Persico, and Todd 2001; Ayres 2002; Persico 2002; Hernández-Murillo and Knowles 2004). If searches of nonwhite drivers are less productive (have a lower likelihood of yielding contraband) then this suggests police might be applying a lower standard of suspicion to nonwhite drivers. More recently, Grogger and Ridgeway (2006) used changes to and from Daylight Savings Time to detect whether the ability to identify race in advance of the stop influenced the race distribution of drivers that were stopped.

All of these methods focus on the police departments as a whole and report whether there is or is not evidence of racially biased policing. Some police executives have suggested that if there are problems then they stem from ''a few bad apples.'' These sentiments are consistent with prior research that has found that a small fraction of police officers in a given department contribute to a disproportionate share of cases of abuse of authority (Sherman 1978).

If racial bias is the result of a few ''problem officers'' then the previously discussed methods that examine bias at the departmental level will not be likely to detect the problem and, even if somehow they have the statistical power to detect the problem, they cannot help to identify potential problem officers. Walker (2001, 2002, 2003b) conceptualized an ''internal benchmark'' method that compares officers' stop decisions with decisions made by other officers patrolling the same area at the same time. The comparison of decisions made in the same areas and times is critical. Research notes that police officer behavior varies as a function of location and time of the day and that those officers working in ''troubled'' areas with heavier workloads become more vigilant about intervening when a situation appears suspicious (Klinger 1997).

This basic internal benchmark strategy has been adopted as a part of several ''early warning systems'' (Walker 2003a). At the

Greg Ridgeway is Senior Statistician, RAND Statistics Group, RAND, 1776 Main Street, Santa Monica, CA 90407 (E-mail: *gregr@rand.org*). John M. MacDonald is Assistant Professor, Department of Criminology, University of Pennsylvania, 483 McNeil Building, 3718 Locust Walk, Philadelphia, PA 19104-6286 (E-mail: *johnmm@sas.upenn.edu*). Partial support for this research came from the New York City Police Foundation. Responsibility for all the content of this article lies with the authors. The authors thank the editor, associate editor, and referees for their comments on earlier versions of this article. The authors also thank Dan McCaffrey, Paul Heaton, Lorie Fridell, and the members of the RAND Statistics Group who reviewed earlier versions of this work and helped refine the methodology.

Journal of the American Statistical Association
June 2009, Vol. 104, No. 486, Applications and Case Studies
DOI 10.1198/jasa.2009.0034

Los Angeles Police Department (LAPD), the TEAMS II Risk Management Information System places officers in one of 33 peer groups (Birotte 2007). Officers in the same peer group presumably are expected to conduct similar policing activities. If an officer exceeds certain thresholds compared with a peer group, such as being in the top 1% on number of complaints or number of use-of-force incidents, the system generates an ''action item'' for follow-up. However, officer roles in LAPD are certainly more diverse than 33 groups can capture and the system generates more action items than reasonably can be investigated. For example, an estimated 16% of the action items occurred after an officer had a single complaint or a single use-of-force incident. The risk of false positives seems high. Similar problems are likely in other audit systems, like Pittsburgh's Performance Assessment and Review System, Cincinnati's Risk Management System, and Phoenix's Personnel Assessment System, which compute a ''peer-officer-based formula'' to flag officers (Walker 2003a), but do not take into account the different environments where officers in the same peer group work.

This article describes a method for constructing a customized internal benchmark for each officer, comparing the race distribution of suspects stopped by the officer in question with the race distribution of suspects stopped by other officers at the same times, places, and contexts. Rather than forming peer groups, this method creates a unique set of comparison stops for each officer, customized to the individual officer's unique assignment and patrol patterns. Our internal benchmark analysis blends three statistical methodologies that are the focus of active research efforts: propensity score weighting, doubly robust estimation, and false discovery rates. We use propensity score weighting to construct each officer's internal benchmark, doubly robust estimation to remove any residual bias and reduce variance, and a false discovery rate analysis to flag potential problem officers. We apply the internal benchmarking methodology to data on 500,000 pedestrian stops that New York City Police Department (NYPD) officers made in 2006 to flag officers who have anomalous patterns. The method flags 15 NYPD officers who appear to be stopping an unusually large fraction of nonwhite pedestrians and flags 13 officers who appear to be stopping substantially fewer nonwhite pedestrians than expected. We show how these three contemporary statistical methods present compelling evidence for that conclusion.

2. DATA, METHODS, AND ANALYSIS

In February 2007, the NYPD released statistics indicating that approximately 500,000 pedestrians had been stopped on suspicion of a crime in New York City in 2006. Almost 90% of the stops involved nonwhites. The number of stops and the apparent lopsided representation of nonwhites among those stopped generated concerns about a systematic pattern of racially biased police practices (New York Civil Liberties Union 2007). Ridgeway (2007a) showed that time, place, and context of the stops could explain much of the racial disparity but also found unexplainable disparities in some areas. The method presented in this article addresses one aspect of this question, namely whether the data suggest that specific individual officers are stopping a disproportionate number of nonwhite pedestrians relative to similarly situated stops made by other officers.

The fundamental goal of internal benchmarking is to compare stops made by a particular officer with stops made by other officers occurring at the same times, places, and contexts. The latter stops form the officer's internal benchmark. The officer's stops and the benchmark stops can be compared on features such as the percentage involving nonwhite suspects or the percentage involving some use of force. Accounting for time, place, and context in the benchmark construction assures us that both the officer and the benchmark are exposed to similar sets of offenses and offenders. In our analysis, we account for month, day of week, time of day, precinct, (x,y) coordinates of the stop location, whether it was a transit or public-housing location, the officer's assigned command, whether the officer was in uniform, and whether the stop was a result of a radio run (i.e., officers were dispatched to a location in response to a report or emergency call).

2.1 Propensity Score Weighting

We match the joint distribution of the features of stops made by other officers to the distribution of features of stops made by the officer in question. Note that we are not matching individual stops to one another, but rather matching the joint distribution of their features. We construct our internal benchmark by reweighting the stops of potential benchmark stops so that the joint distributions align. Specifically,

$$f(\mathbf{x}\,|\,t=1) = w(\mathbf{x})f(\mathbf{x}\,|\,t=0) \qquad (1)$$

where $\mathbf{x}$ is the vector of stop features, t is a 0/1 indicator for a stop involving the officer under examination, and $w(\mathbf{x})$ is the weight function, for which we solve to equalize the feature distributions. Solving for $w(\mathbf{x})$ and applying Bayes's theorem to the two conditional distributions of $\mathbf{x}$ yields

$$w(\mathbf{x}) = \frac{f(t=1\,|\,\mathbf{x})}{f(t=0\,|\,\mathbf{x})}K \qquad (2)$$

where K is a constant that does not depend on $\mathbf{x}$ and will cancel in the outcomes analyses. The probability that a stop having features $\mathbf{x}$ involves the officer in question, $f(t=1|\mathbf{x})$, is the propensity score (Rosenbaum and Rubin 1983). In traditional propensity score analysis t is the treatment indicator; here, stops made by the officer in question are deemed to be exposed to the ''treatment'' and stops made by other officers are the ''control'' stops.

We will denote the propensity score for stop i as p_i. According to (2), weighting the stops of other officers by $p_i/(1-p_i)$ will align the distribution of all of their stop characteristics with the distribution of the target officer's stop characteristics. Those stops having features (time, location, context) that are quite different from the characteristics of stops that the target officer makes will have propensity scores near 0 and therefore will receive weights near 0. Stops with large propensity scores, on the other hand, have features that are very similar to the target officer's stops and will have larger weights. These contribute most to the target officer's internal benchmark. For more detailed analysis of propensity score weights see Hirano and Imbens (2001), Wooldridge (2001, pp. 614–621), McCaffrey, Ridgeway, and Morral (2004), and Ridgeway (2006).

We use the boosted logistic regression method described in McCaffrey et al. (2004) to compute the propensity scores. This method essentially estimates the propensity scores from a logistic regression model constrained with an L_1 penalty on the size of the coefficients (Tibshirani 1995). The associated penalized log-likelihood function is

$$\ell(\boldsymbol{\alpha}) = \sum_{i=1}^{n} t_i \boldsymbol{\alpha}'\mathbf{h}(\mathbf{x}_i) - \log\left(1 + e^{\alpha'\mathbf{h}(\mathbf{x}_i)}\right) - \lambda \sum_{j=1}^{J} |\alpha_j| \qquad (3)$$

where $\mathbf{h}(\mathbf{x})$ is some suitable class of basis functions. The second summation on the right side of the equation is a penalty term that decreases $\ell(\boldsymbol{\alpha})$ when there are coefficients that are large in absolute value. Setting $\lambda = 0$ returns the standard (and potentially unstable) logistic regression estimates of $\boldsymbol{\alpha}$. Setting λ to be very large essentially forces all of the α_j to be near 0 (the penalty excludes α_0). For a fixed value of λ the estimated $\hat{\boldsymbol{\alpha}}$ can have many coefficients exactly equal to 0, not just extremely small but precisely 0, and only the most powerful predictors of t will be nonzero. As a result the absolute penalty operates as a coefficient shrinkage and variable selection penalty. In practice, if we have several predictors of t that are highly correlated with each other, (3) tends to include several of them in the model, shrinks their coefficients toward 0, and produces a predictive model that utilizes all of the information in the covariates, producing a model with greater out-of-sample predictive performance than models fit using variable subset selection methods.

We let $\mathbf{h}(\mathbf{x})$ be a large collection of piecewise constant functions of the x_j variables and their interactions. That is, in $\mathbf{h}(\mathbf{x})$ we include indicator functions like I(month = January), I(location = transit), and interactions among them like I(month = January) × I(location = transit). This collection of basis functions spans a plausible set of propensity score functions, is computationally efficient, and is flat at the extremes of $\mathbf{x}$, thereby reducing the risk that the propensity score estimates are spuriously near 0 and 1 (which can occur with linear basis functions of $\mathbf{x}$). Theoretically we can estimate the model in (3), selecting a λ small enough so that it will eliminate most of the irrelevant terms and yield a parsimonious model with only the most important main effects and interactions. Boosting (Friedman 2001) effectively implements this strategy using a computationally efficient method that Efron, Hastie, Johnstone, and Tibshirani (2004) showed is equivalent to optimizing (3). We used the implementation in the generalized boosted modeling package in R (Ridgeway 2007b). Whereas boosting has the potential to underfit or overfit, selecting λ using cross-validation or to optimize balance between the target officer's stops and the benchmark stops, as we do, minimizes those risks. Mease, Wyner, and Buja (2007) challenge the quality of class probabilities computed from boosted logistic regression models. However, they do not regularize the models or use effective stopping rules like cross-validation. In constructing the benchmark for each officer, we selected λ so that the resulting propensity score weights produced weighted marginal feature distributions for the benchmark stops that matched the marginal feature distributions for the officer's stops.

As an example, Table 1 presents the internal benchmark calculated for one particular NYPD officer. This officer made 392 stops in 2006. The largest fraction of stops occurred on Thursdays, mostly in precinct B, and mostly during the night shift. This officer made few stops in January and never made stops in public housing or transit locations. This officer's stops involved black pedestrians 83% of the time. To assess whether 83% is a reasonable fraction we construct a weighted set of stops for the benchmark.

The ''Internal Benchmark'' column in Table 1 shows the marginal distributions of features for stops made by other officers weighted as in (2). The benchmark stops have almost

Table 1. Illustration of internal benchmarking for an example officer

		Example	Internal
Stop		Officer (%)	Benchmark (%)
Characteristic		(n = 392)	(ESS = 3, 676)
Month	January	3	3
	February	4	4
	March	8	9
	April	7	5
	May	12	12
	June	9	9
	July	7	7
	August	8	9
	September	10	10
	October	11	10
	November	11	11
	December	9	10
Day of the week	Monday	13	13
	Tuesday	11	10
	Wednesday	14	15
	Thursday	22	21
	Friday	15	16
	Saturday	10	11
	Sunday	15	14
Time of day	12–2 a.m.	11	11
	2–4 a.m.	5	5
	10 a.m.−12 p.m.	0	1
	12–2 p.m.	12	13
	2–4 p.m.	13	12
	4–6 p.m.	9	10
	6–8 p.m.	8	8
	8–10 p.m.	23	23
	10 p.m.−12 a.m.	17	17
Patrol borough	Brooklyn North	100	100
Precinct	A	0	0
	B	98	98
	C	1	1
	D	1	0
Inside or outside	Inside	4	6
	Outside	96	94
Housing or transit	Transit	0	0
	Housing	0	0
	Other	100	100
Command assignment	Precinct B	100	100
In uniform	Yes	99	97
Radio run	Yes	1	3

NOTE: The precincts have been given random letter codes to mask the officer's identity. For the benchmark stops ESS represents the effective sample size, $ESS = (\Sigma w_i)^2/\Sigma w_i^2$.

exactly the same distribution of features as the target officers; they were made in the same places, times, and contexts. For our purposes, the fact that we can balance the stop features as in Table 1 is sufficient to show our method's effectiveness.

Whereas Table 1 shows that the method matches univariate marginal distributions for discrete stop features, distributions of continuous stop features and higher dimensional marginal distributions are also well matched. For example, we included the latitude and longitude of the stops in the propensity score model in addition to the discrete stop features listed in Table 1. Figure 1 shows a map of NYPD precinct B with a few additional sectors of the adjacent precincts where the officer in question made stops. The contours in Figure 1 show the distributions of the stop locations for the officer's stops (left panel) and the benchmark stops (right panel). This demonstrates that not only the officer's stops and the benchmark match in terms of the percentage of stops in precinct B (as shown in Table 1), but also the benchmark is further customized for the specific parts of the precinct in which the officer patrols. The scale of the map and the locations of the stops suggest that this officer was on a foot patrol.

2.2 Outcome Analysis With a Propensity Score Weighted Benchmark

Whereas 83% of the officer's stops involved black pedestrians, the weighted benchmark stops involved black pedestrians 78% of the time. We need to resolve when these differences from the benchmark warrant closer scrutiny. Several of the early warning systems flag officers based on z scores, which have been promoted as a measure for flagging problem officers such as those officers exceeding standard significance levels. Fridell (2004) suggests 2.0 and Smith (2005) suggests 1.645. We argue later that such cutoffs generate too many false positives to be useful.

The z-score for our proposed benchmark constructed using propensity score weights can be obtained as $z = \hat{\beta}_1/\mathrm{se}(\hat{\beta}_1)$ from maximizing the weighted logistic regression log-likelihood

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} w_i\big(y_i(\beta_0 + \beta_1 t_i) - \log\big(1 + e^{\beta_0+\beta_1 t_i}\big)\big) \qquad (4)$$

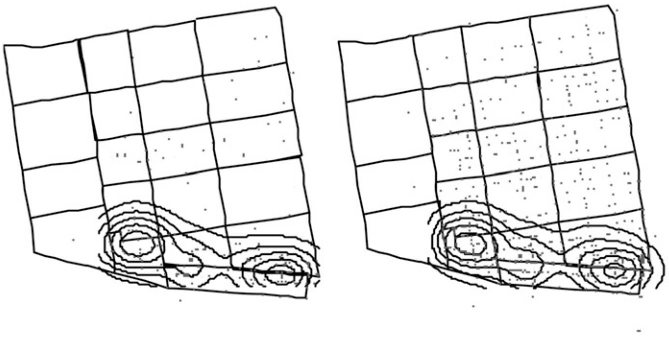

Figure 1. Maps of the locations of the officer's stops (left) and the locations of the benchmark stops (right). The contours show the region that has the highest concentration of stops. This demonstrates that the propensity score based benchmark can also match on continuous stop features and on higher dimensional marginal distributions.

where y_i is the 0/1 indicator of whether the stop involved, for example, a black pedestrian. Another outcome of interest could be used for y_i, such as whether the stop involved a member of another racial group, involved use of force, resulted in a complaint, or resulted in a commendation. The target officer's stops receive $w_i = 1$ and the stops made by other officers receive $w_i = p_i/(1 - p_i)$ as in (2). The standard error of β_1 can be estimated with a sandwich estimator to account for the weights (Huber 1967). For the officer described in Table 1, $z = 2.4$.

If the propensity score weights are effective at equating the distribution of the officer's stop features, $f(\mathbf{x}|t = 1)$, with the distribution of the benchmark stop features, $w(\mathbf{x})f(\mathbf{x}|t = 0)$, then the z statistics will not contain the potential confounding effect of $\mathbf{x}$. If differences remain between $f(\mathbf{x}|t = 1)$ and $w(\mathbf{x})f(\mathbf{x}|t = 0)$ then additional regression adjustment can be effective at removing small differences and producing doubly robust estimates.

2.3 Doubly Robust Estimation

Achieving balance in the feature distributions of the officer's stops and the benchmark stops is critical to a fair internal benchmark analysis. For most officers the propensity score weights construct a convincing comparison set of stops, although for 10% of the officers at least one stop feature differed from the benchmark stops by more than 5%; the largest difference was 12%. If those features on which the officer's stops and benchmark stops differ are strongly associated with the suspect's race, then the analysis can be biased. Doubly robust methods (Bang and Robins 2005; Kang and Schafer 2007) for estimating effects can reduce the risk of bias. We obtain a doubly robust estimate of the benchmark minority stop rate by expanding (4) to include covariates. We fit a weighted logistic regression model, obtaining $\boldsymbol{\beta}$ that maximizes the propensity score weighted log-likelihood

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} w_i\Big(y_i s(t_i, \mathbf{x}_i|\beta) - \log\Big(1 + e^{s(t_i,\mathbf{x}_i|\beta)}\Big)\Big) \qquad (5)$$

where $s(t,\mathbf{x}|\beta)$ is a regression function for which we use a linear model, $\beta_0 + \beta_1 t + \gamma'\mathbf{x}$, but more flexible options are possible. If *either* the propensity score weights accurately match the officer and benchmark stop feature distributions *or* the regression model, s, correctly specifies the relationship between $(t, \mathbf{x})$ and $\log P(y = 1|t, \mathbf{x})/P(y = 0|t, \mathbf{x})$, then the estimator

$$\bar{y}_0^{DR} = \sum_{i=1}^{n} t_i \frac{1}{1 + e^{-s(0,\mathbf{x}_i|\hat{\beta})}} \qquad (6)$$

is a consistent estimator of the benchmark minority stop rate (the percentage of the officer's stops that we would have expected to have been minorities based on the minority stop rate of other officers patrolling under the same conditions).

The expression in (6) computes what is commonly referred to as the "recycled prediction" estimate of the benchmark outcome. The model in (5) can be fit using a weighted logistic regression program that treats the w_i as survey weights (e.g., R's survey package, Stata's svy: logit, SAS's PROC SURVEYREG). If $s(t,\mathbf{x}|\boldsymbol{\beta})$ has no interactions between t and $\mathbf{x}$ then the z-statistic for β_1 is an appropriate measure of the difference between the officer's stops and the benchmark, though the next section indicates that $N(0, 1)$ might not be the appropriate reference distribution.

If $s(t,\mathbf{x}|\boldsymbol{\beta})$ includes interactions between t and $\mathbf{x}$, then we would estimate the officer's departure from the benchmark with the regression adjusted difference as

$$\hat{\theta} = \sum_{i=1}^{n} t_i \left(\frac{1}{1+\exp(-s(1,\mathbf{x}_i|\hat{\boldsymbol{\beta}}))} - \frac{1}{1+\exp(-s(0,\mathbf{x}_i|\hat{\boldsymbol{\beta}}))} \right). \tag{7}$$

The z-statistic would then be $\hat{\theta}/\mathrm{se}(\hat{\theta})$. To compute $\mathrm{se}(\hat{\theta})$ we would use the subset of the data containing just the officer's stops (stops with $t = 1$). We would append to that dataset a replicate of the officer's stops except setting $t = 0$. With the fitted logistic regression model we would generate predictions for this dataset with $2\sum t_i$ observations, storing the covariance matrix of the predictions as $\boldsymbol{\Sigma}$. We would then estimate $\mathrm{Var}(\hat{\theta})$ as $\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}$ where $\mathbf{a}$ has length $2\sum t_i$ with the first $\sum t_i$ elements equal to 1 and the last $\sum t_i$ elements equal to – 1. For the officer described in Table 1, the regression adjustment yields $z = 2.2$, slightly less than the z-score computed without regression adjustment (2.4).

In 2006, 15,855 officers completed at least one stop. Of those officers, 3,034 officers had more than 50 pedestrian stops. We constructed benchmarks only for those officers who made at least 50 stops in 2006, focusing the analysis on those officers most frequently involved in pedestrian stops. This restriction provides a minimum level of statistical power for detecting differences if they exist. For each officer, in turn, we constructed a separate internal benchmark like the one shown in Table 1, one specifically tailored for each officer's patterns of stops. For 278 of those officers, a suitable set of benchmark stops could not be constructed. For these officers, the best set of benchmark stops differed from the given officer's stops by more than 10% on some observable factors. These officers generally made fewer stops and had uncommon assignments (e.g., not in uniform, making stops in precincts in which few other officers make stops). Our final set for analysis includes 2,756 officers. These officers represent 7% of NYPD officers and 17% of those involved in stops, but collectively they made 54% of all the stops for the NYPD. After computing the 2,756 z scores, we compute for each officer the probability that, given their z-score, the officer's stop pattern differs from the benchmark.

2.4 False Discovery Rates

An officer's z-statistic is a measure of the degree to which the racial distribution of the officer's stops differs from those of the officer's benchmark. Issues of multiple testing clearly arise when attempting to infer statistical differences based on 2,756 z-statistics. Benjamini and Hochberg (1995) pioneered the use of the false discovery rate (fdr) as an alternative multiple comparison adjustment technique. The fdr is the probability of no difference between the officer and the benchmark given the value of an observed test statistic, z. Our method flags those officers who have values of z that suggest a high probability of exceeding their benchmark.

If one conceptualizes that an officer is either "problematic" (the racial distribution of the officer's stops does not match that of the corresponding benchmark) or "not problematic" (the racial distribution of the officer's stops matches that of the benchmark), then one can derive the probability of an officer being problematic as

$$\begin{aligned} P(\text{problem}|z) &= 1 - P(\text{no problem}|z) \\ &= 1 - \frac{f(z|\text{no problem})f(\text{no problem})}{f(z)} \\ &\geq 1 - \frac{f_0(z)}{f(z)} \end{aligned} \tag{8}$$

where $f_0(z)$ is the distribution of z for nonproblem officers and $f(z)$ is the distribution of z for all officers (Efron 2004). The expression in (8) is 1 – fdr. If most officers are not problem officers (e.g., $f(\text{no problem}) > 0.90$) then the bound in (8) is near equality.

Figure 2 shows a histogram of the 2,756 z scores computed for the NYPD data. In standard circumstances, $f_0(z)$ would be an $N(0, 1)$ density. However, in a collection of 2,756 z scores, each of which is correlated with the other, the empirical distribution of the z statistics may be much wider than an $N(0, 1)$ (Efron 2007). As shown in Figure 2, the $N(0, 1)$ overlaid is much narrower than the histogram of the z scores. Following Efron (2004), we estimate $f_0(z)$ with the empirical null assumed to be normal with parameters estimated using the location and curvature of the central part of the histogram, resulting in an $N(0.1, \sigma = 1.4)$ density. We estimate $f(z)$ with a nonparametric density estimate. We used the locfdr R package for these computations (Efron, Turnbull, and Narasimhan 2008).

Officers with $z > 4.2$, representing five of the 2,756 NYPD officers, have probability in excess of 0.50 of having a distribution of stops more extreme than that of the benchmark (e.g., higher proportion of stops of black pedestrians). The choice of 0.50 as a threshold implies that the cost of failing to identify a problem officer equals the cost of flagging a good officer, which may not be the case. We recommend to departments to start with those above 0.50, assess issues related to those officers, and if the process identifies substantive concerns then proceed through the list until a meaningful false-positive risk suggests that it would be appropriate to stop. Table 2 summarizes the five officers with fdr < 0.50. In each case these officers appear to be making a substantially higher fraction of stops of black pedestrians than their benchmark. On

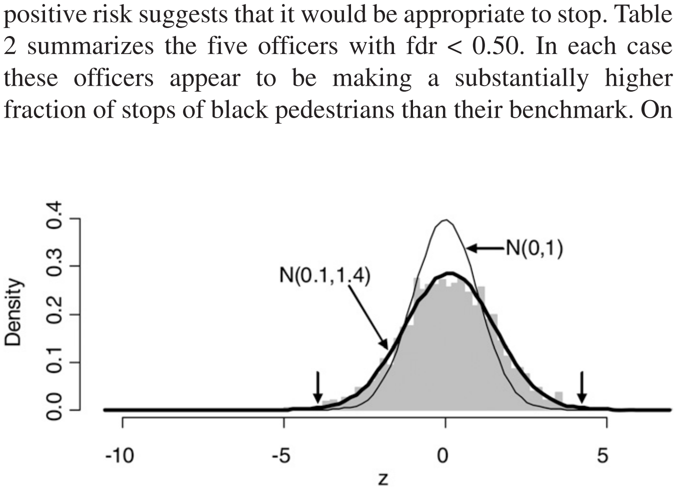

Figure 2. The distribution of the 2,756 z-statistics and the estimated reference distribution. The $N(0.1, 1.4)$ density is the estimated $f_0(z)$, the distribution of the z scores of officers who match their benchmark assuming normality. The downward pointing arrows mark the value of z for which the $P(\text{problem}|z) = 0.5$.

Table 2. Internal benchmark analysis for stop rates of black suspects

Officer		Benchmark		
Black (%)	Stops (n)	Black (%)	Stops (n)	fdr
86	151	55	773	0.03
85	218	67	473	0.38
77	237	56	1,081	0.14
75	178	51	483	0.22
64	59	20	695	0.02

the left side of the distribution we also find 12 officers with $z < -4.0$ who stopped many fewer black suspects than their benchmarks suggest. Repeating the analysis for stops of Hispanic pedestrians identifies 10 officers who appear to be overstopping and four who appear to be understopping Hispanic pedestrians.

The procedure currently being advocated to police departments of flagging officers with $z > 2.0$ selects 9% of the officers and the $z > 1.645$ selects 13% of the officers. Of the 242 officers with $z > 2.0$, 217 (90%) of them have fdr estimated to be greater than 0.999, indicating that they are unlikely to truly deviate from their benchmark. This analysis provides clear evidence that standard auditing methods create too many false positives to be useful.

3. LIMITATIONS OF THE ANALYSIS

There are several limitations and cautions warranted when utilizing the proposed internal benchmark for flagging officers. Omitted variable bias is possible in all studies using observational data. If there is a confounding variable (besides racial bias) that is associated with both the officer and the likelihood of stopping a nonwhite pedestrian, then the estimated race effect will be biased. The analysis uses all observable features of time, place, and assignment that are clearly confounding variables, but an unmeasured variable may explain the observed differences. There also may be legitimate reasons for some of the observed officer differences, such as an idiosyncratic assignment. Our analysis, however, indicates that the racial differences in stops for the outlying officers observed is not because of their unique exposure to a racial distribution of offenses and offenders by place, time, reason for the stop, or random chance. At this stage we do not know whether these flagged officers are engaged in racially biased policing, but the patterns observed suggest the need to audit these officers' stop decisions.

Implicit in the proposed framework, which draws on a multiple-comparison idea relevant to hypothesis testing, is an assumption that numerous officers have the same level of bias, which is either near zero or identically equal to zero. Although the method compares officers to their peers, it is not necessarily the case that their peers are unbiased. If, for example, all of the officers in a precinct act in a racially biased manner then when each is compared with the others, none of the officers in this precinct will be flagged as problematic. Only in the case that most officers are unbiased and only a few are problematic, the setting several police executives have suggested, will the method actually measure race bias among officers. An alternative conception of the problem, based on assuming a distribution of bias for the officers, might be able to improve upon this framework. Select comparisons might be able to inform us about this distribution. For example, if some of the officers from the highly biased precinct made stops outside of their precinct, we could compare those stops to stops made by officers from the adjacent precinct. This would suggest estimates of the bias for those in the problematic precinct to be relatively larger than for those officers in the adjacent precinct. Further development of such ideas may result in an fdr calculation that deconvolves the race bias distribution from the reference distribution.

An additional limitation to this analysis is the possible practice of systematic under-reporting of stops. If certain officers systematically under- or over-report their stops of nonwhite pedestrians, the estimated race effect will be biased. The NYPD has multiple layers of auditing to ensure that pedestrian stops are documented completely and contain valid and sufficiently detailed entries to each question. The NYPD audit system, however, does not address whether undocumented stops are occurring among specific officers. Because officers have an incentive to demonstrate productivity through stops, most stops should be documented. However, particularly problematic stops may not be recorded. Ridgeway (2007a) recommended a study of radio communications, monitoring them for a fixed period in a few randomly selected precincts, and determining whether stop forms exist that match the times and places of reported street encounters. Ridgeway, Schell, Riley, Turner, and Dixon (2006) audited Cincinnati PD traffic stops, comparing them with dispatch logs and found that completion rates were above 96%.

Our analysis computed benchmark comparisons for only those officers making more than 50 stops. Whereas these officers cover the majority of pedestrian stops, this cutoff prevents the analysis from detecting biases in those officers making fewer than 50 stops. There is limited statistical power to detect differences in minority stop rates when an officer has fewer than 50 stops, and it is conceivable that an officer could simply make fewer stops to underpower his or her benchmark to avoid the possibility of being flagged. Combining data across years could address this problem within the proposed framework. A hierarchical modeling framework that smooths estimates of racial bias could also be considered to accommodate officers with fewer than 50 stops; we would not expect such an approach to flag many additional officers, but it is conceivable that an officer with fewer than 50 stops but a large racial discrepancy from benchmark predictions would stand out using such a method.

Our analysis also dropped 278 officers for whom we could not construct an adequate benchmark. The problem occurs when some officers had very unique assignments. For example, there might be one officer who was essentially the only officer to make stops in a particular public housing complex. As a result the method could not identify similarly situated stops made by other officers. Technically, we could still compute a doubly estimate, but its accuracy would rely heavily on the regression model's ability to extrapolate correctly. For example, if an officer made 70% of his stops in a particular neighborhood but less than 60% of the benchmark consisted of stops in that neighborhood then we are relying heavily on the regression model to adjust for the difference. We, as well as police managers we have worked with, are skeptical of relying on regression models too heavily when there are variables (such as neighborhood) that, even after weighting, are strongly associated with both officer and pedestrian race. We selected 10% as the cutoff to drop officers because we felt it generated a certain level of suspicion among police management that the benchmark was inadequate and not useful. Again, combining data across years or more sophisticated regression models may be able to relax this restriction.

Further, the estimate of $f_0(z)$, the distribution of z for non-problem officers, requires a large number of officers. NYPD has a sufficient number to estimate this reasonably well, but without strong assumptions on $f_0(z)$, an internal benchmark analysis for smaller departments could not yield reliable estimates of the probability of being a problematic officer.

4. DISCUSSION

This study was predicated on the notion that methodological approaches for detecting racial bias in police decision-making need to be sensitive to the environmental context in which officers carry out their daily work and make stops. Prior research suggested that officers should behave differently depending on the location in which they are working. The internal benchmark method controlled for this potential confounding factor on police officers' decisions to stop pedestrians. It compares officers patrolling the same areas at the same times, assuming that conditioning on these factors would result in officers being exposed to the same population of offenders. If the officers all had the same duties, then we would expect the race distribution of their stops to be similar, if not the same. After using a doubly robust benchmark construction to compare the racial distribution of the stops of 2,756 officers, we found five officers who appeared to be stopping a significantly larger fraction of black pedestrians and 10 officers stopping an excessive fraction of Hispanic pedestrians when compared with stops other officers made at the same times and places.

The results from this study suggest that "problematic officers" might be apparent when accounting for the environmental context in which officers perform their stops. Importantly, the differences observed in the current study are not likely the result of chance differences or differences in the exposure to unique times or locations in New York City. This analysis gives police management useful tools to identify potential problem officers whose work performance may need closer inspection. As opposed to officers monitored with other risk management systems currently in operation, the proposed methodology has a much lower risk of generating "false positives," subjecting some officers to unnecessary scrutiny and potentially concealing problem officers among large numbers of false positives.

Whereas this analysis is limited to within-agency variation in racial differences in pedestrian stop decisions, we think this is an important avenue for future legal and empirical inquiry on racial bias in police decision-making. The internal benchmark approach used here moves the debate away from discussions about the relevant aggregate benchmarks and focuses on providing a rigorous method for comparing officers to each other. This methodology could also be expanded to evaluate the volume of complaints, the use of arrest authority, and other aspects of the legal decision-making for police officers.

Software implementing this method was deployed at the Cincinnati PD in 2007 and is now run quarterly as a standard part of every patrol officer's review. Accordingly, we expect there would be substantial public-policy interest in future evaluations of this program, which we hope will be forthcoming.

[Received July 2008. Revised November 2008.]

REFERENCES

Alpert, G., Smith, M., and Dunham, R. (2004), "Toward a Better Benchmark: Assessing the Utility of Not-at-Fault Traffic Crash Data in Racial Profiling Research," *Justice Research and Policy,* 6(2), 44–69.

Ayres, I. (2002), "Outcome Tests of Racial Disparities in Police Practices," *Justice Research and Policy,* 4(1), 131–142.

Bang, H., and Robins, J. (2005), "Doubly Robust Estimation in Missing Data and Causal Inference Models." *Biometrics* 61, 962–972. Corrected in 2008, *Biometrics,* 64, 650.

Benjamini, Y., and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society: Series B,* 57, 289–300.

Birotte, A. (2007, November), "Training Evaluation and Management System (TEAMS) II Audit, Phase I (Fiscal Year 2007/2008)," Technical Report, Office of the Inspector General, Los Angeles Police Department, Los Angeles, CA. Available at *http://www.lacity.org/oig/Reports/TEAMS2F1Report_11-06-07.pdf*.

Efron, B. (2004), "Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis," *Journal of the American Statistical Association,* 99, 96–104.

———(2007), "Correlation and Large-Scale Simultaneous Significance Testing," *Journal of the American Statistical Association,* 102 (477), 93–103.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," *Annals of Statistics,* 32(2), 407–499.

Efron, B., Turnbull, B., and Narasimhan, B. (2008), *locfdr 1.1 Package Manual.* R project.

Fagan, J., and Davies, G. (2000), "Street Stops and Broken Windows: Terry, Race and Disorder in New York City," *The Fordham Urban Law Journal,* 28, 457–504.

Fridell, L. A. (2004), *By the Numbers: A Guide for Analyzing Race Data from Vehicle Stops*, Washington, DC: Police Executive Research Forum.

Friedman, J. (2001), "Greedy Function Approximation: A Gradient Boosting Machine," *The Annals of Statistics,* 29(5), 1189–1232.

Gelman, A., Fagan, J., and Kiss, A. (2007), "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, 813–823.

Grogger, J., and Ridgeway, G. (2006), "Testing for Racial Profiling in Traffic Stops From Behind a Veil of Darkness," *Journal of the American Statistical Association,* 101(475), 878–887.

Hernández-Murillo, R., and Knowles, J. (2004), "Racial Profiling or Racist Policing? Testing in Aggregated Data," *International Economic Review,* 45(3), 959–989.

Hirano, K., and Imbens, G. (2001), "Estimation of Causal Effects Using Propensity Score Weighting: An Application to Data on Right Heart Catheterization," *Health Services and Outcomes Research Methodology,* 2, 259–278.

Huber, P.J. (1967), "The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Volume I, Berkeley: University of California Press, pp. 221–233.

Kang, J., and Schafer, J. (2007), "Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean From Incomplete Data," *Statistical Science,* 22(4), 523–580.

Kennedy, R. (1997), *Race, Crime, and the Law*, New York: Pantheon Books.

Klinger, D. A. (1997), "Negotiating Order in Patrol Work: An Ecological Theory of Police Response to Deviance," *Criminology,* 35(2), 277–306.

Knowles, J., Persico, N., and Todd, P. (2001), "Racial Bias in Motor Vehicle Searches: Theory and Evidence," *The Journal of Political Economy,* 109, 203–229.

Lamberth, J. (1994), "Revised Statistical Analysis of the Incidence of Police Stops and Arrests of Black Drivers/Travelers on the New Jersey Turnpike Between Exits or Interchanges 1 and 3 From the Years 1988 Through 1991," Technical Report, Temple University, Department of Psychology.

McConnell, E. H., and Scheidegger, A. R. (2001), "Race and Speeding Citations: Comparing Speeding Citations Issued by Air Traffic Officers With

Those Issued by Ground Traffic Officers," in *Annual Meeting of the Academy of Criminal Justice Sciences*, Washington, D.C.

McCaffrey, D., Ridgeway, G., and Morral, A. (2004), "Propensity Score Estimation With Boosted Regression for Evaluating Causal Effects in Observational Studies," *Psychological Methods,* 9(4), 403–425.

Mease, D., Wyner, A., and Buja, A. (2007), "Boosted Classification Trees and Class Probability/Quantile Estimation," *Journal of Machine Learning Research,* 8, 409–439.

Montgomery County Department of Police (2002), *Traffic Stop Data Collection Analysis*, (3rd report), Montgomery County Department of Police.

National Research Council (2003), *Fairness and Effectiveness in Policing: The Evidence*, Washington, DC: The National Academies Press.

New York Civil Liberties Union (2007, February), "Long-Awaited "Stop-and-Frisk"Data Raises Questions About Racial Profiling and Overly Aggressive Policing, NYCLU says,"Press Release, Available at http://www.nyclu.org/node/919..

Persico, N. (2002), "Racial Profiling, Fairness, and Effectiveness of Policing," *The American Economic Review,* 92(5), 1472–1497.

Ridgeway, G. (2006), "Assessing the Effect of Race Bias in Post-Traffic Stop Outcomes Using Propensity Scores," *Journal of Quantitative Criminology,* 22(1), 1–26.

———(2007a), "Analysis of Racial Disparities in the New York Police Department's Stop, Question, and Frisk Practices," Technical Report TR-534-NYCPF, RAND Corporation.

———(2007b), *GBM 1.6-3 Package Manual*, R Project.

Ridgeway, G., Schell, T. L., Riley, K. J., Turner, S., and Dixon, T. L. (2006) "Police-Community Relations in Cincinnati: Year Two Evaluation Report," Technical Report TR-445-CC, RAND Corporation, Santa Monica, CA. Available at http://www.rand.org/pubs/technical_reports/TR445/.

Rosenbaum, P., and Rubin, D. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika,* 70, 41–55.

Russell-Brown, K. (1998), *The Color of Crime: Racial Hoaxes, White Fear, Black Protectionism, Police Harassment and Other Macro-Aggressions*, New York: New York University Press.

Sherman, L. W. (1978), *Scandal and Reform: Controlling Police Corruption*, Berkeley: University of California Press.

Smith, M. R. (2005), "Depoliticizing Racial Profiling: Suggestions for the Limited Use and Management of Race in Police Decision-Making," *George Mason University Civil Rights Law Journal,* 15(2), 219–260.

Steward, D. (2004), Racial Profiling: Texas Traffic Stops and Searches, Technical Report, Steward Research Group, Austin, TX.

Tibshirani, R. (1995), "Regression Selection and Shrinkage Via the Lasso," *Journal of the Royal Statistical Society: Series B,* 57, 267–288.

Walker, S. (2001), "Searching for the Denominator: Problems with Police Traffic Stop Data and an Early Warning System Solution," *Justice Research and Policy,* 3(2), 63–95.

———(2002), "The Citizen's Guide to Interpreting Traffic Stop Data: Unraveling the Racial Profiling Controversy," unpublished manuscript.

———(2003a, "Early Intervention Systems for Law Enforcement Agencies: A Planning and Management Guide," Technical Report, Office of Community Oriented Policing Services, U.S. Department of Justice, Washington DC. Available at http://www.cops.usdoj.gov/html/cd_rom/inaction1/pubs/EarlyInterventionSystemsLawEnforcement.pdf.

———(2003b), "Internal Benchmarking for Traffic Stop Data: An Early Intervention System Approach," Technical Report, Police Executive Research Forum.

Weiss, A., and Grumet-Morris, A. (2005, July), "Illinois Traffic Stop Statistics Act Report for the Year 2004," Technical Report, Northwestern University Center for Public Safety, Evanston, IL. Available at http://www.dot.state.il.us/trafficstop/2004summary.pdf.

Wooldridge, J. (2001), *Econometric Analysis of Cross Section and Panel Data*, Cambridge: MIT Press.