Detecting Inconsistency in Governance[*]

Ryan Copus[†] and Ryan Hübert[‡]

July 26, 2018

Abstract

Measurements of inconsistency among decision makers are important for assessing the quality of governance, as well evaluating policy reforms. And yet, high quality measures of inconsistency in real-world institutions are rarely available. The core problem is that mean differences on outcomes between decision makers (disparity statistics) systematically understate inconsistency. Our contribution is twofold. First, we show how this downward bias can theoretically be eliminated by targeting a specific kind of heterogeneous treatment effect. Second, we demonstrate how machine learning can be used to optimally implement the theory on observational data. We leverage an original dataset of civil appeals in the Ninth Circuit to provide one of the first high quality measures of decision making inconsistency in the court.

```
Abstract:       118 words
Body:         8,213 words
References:     635 words
```

Supplementary material for this article is available online.

[*] We thank participants at the 2016 meeting of the Midwest Political Science Association, the 2016 Political Economy and Public Law conference, the 2017 meeting of the Southern Political Science Association and the 2017 meeting of the American Political Science Association, as well as Tom Clark, Morgan Hazelton, Alan Kluegel, Joy Milligan, Kevin Quinn, Amy Semet, and Aaron Smyth for helpful comments. All errors and omissions are our own.

[†] Ryan Copus is a Climenko Fellow at Harvard Law School. Email: rwcopus@gmail.com.

[‡] Ryan Hübert is Assistant Professor of Political Science at the University of California, Davis. Email: rhubert@ucdavis.edu. *Corresponding author.*

1 Introduction

Discretion to a potentially large set of semi-autonomous administrators and judges is a fact of life for any government. Policy makers at the top levels of government may articulate policies and priorities, but it is these lower-level officials who make those policies and priorities into reality (Lipsky 1980; Ho 2017). We refer to systems in which administrators or judges operate as *decentralized adjudication systems*. Decentralized adjudication systems are distinguished by their routine adjudicative, administrative or enforcement functions. Policy makers may try to adjust policy to constrain the behavior of those administrators and judges, but even then, there are limits to their ability to do so. Sometimes, these officials have different preferences than their political superiors and seek to selectively enforce policy, generating policy drift. But policy drift is not the sole problem confronting political principals; they are also concerned with the *quality* of administration and enforcement.

One important element of quality is the degree to which administrators and judges make decisions consistently. Thus, measurements of inconsistency are important for political principals who seek to advance their policy goals, as well as the public more broadly. For example, high-profile fatal incidents at nursing homes have raised awareness of the variation in regulatory enforcement across states, including variation in the size of fines levied for violations. As a result, the U.S. Centers for Medicare and Medicaid Services has worked to reduce inter-state variation in nursing homes inspections (Ornstein and Groeger 2012). Municipal governments often employ large cadres of inspectors and enforcement officers to monitor restaurants, to issue and verify building permits, to ensure sidewalks are clear of snow, to fine parking violators, etc. The degree of consistency is often a concern. A 2011 grand jury oversight report on the Building Services Division in Oakland, California found (emphasis added):

> [C]ode enforcement inspectors have aggressively pursued blight and sub-standard prop-

> erties throughout Oakland as determined by their *individual interpretations* of the applicable city code. This has led to an *inconsistent enforcement program* backed by inspectors' threats of filing large liens on the offending properties.
>
> Alameda County Grand Jury (2011)

In the context of immigration, consular and border officials all around the world process a large caseload of applications from non-citizens seeking to enter the United States. High profile denials spark renewed attention to the seeming arbitrariness of visa or entry decisions (see for example, Turnbull 2013; Narea 2017). For example, a recent Supreme Court decision permitted the Trump administration to deny entry to immigrants covered by Executive Order 13780 (the "Travel Ban") as long as those immigrants have no "bone fide relationship" with a U.S. person or entity (*Trump v. Int'l Refugee Assistance Project* 2017). Such determinations are made on a case by case basis by immigration officials. A policy maker overseeing immigration officials may observe that those officials deny entry to 80% of immigrants covered by the Travel Ban. On its face, this number provides some information about the overall degree of laxity of U.S. immigration officials. However, if a policy maker's goals are more sophisticated—for example, that the U.S. should deny entry to specific kinds of *high risk* immigrants—then this mean provides limited information. Indeed, if the policy maker *also* knew that, on average, immigration officers make different decisions on 40% of of all applications, then the policy maker might conclude that the agents are not receiving sufficient guidance on how to identify high risk immigrants.

What might a policy maker do with this information? At the most basic level, it can serve as a gut check for how well an institution or organization is performing. Indeed, recent work has focused attention to the potential costs of inconsistency and implored institutions to take it seriously. Kahneman et al. (2016), for example, argues that "noise audits" can alert institutions to surprisingly high and intolerable levels of inconsistency. But beyond the basic goal of understanding the degree of inconsistency, a policy maker may also be interested in *reducing* inconsistency among decision makers. A high profile example comes from the judiciary, where Congress has debated for decades about the quality of decision making in the Ninth Circuit. Scholars, judges, and politicians have contended that the Ninth Circuit is chaotic, too inconsistent, and the home of "jackpot justice." These accusations have led some to argue that splitting the Ninth Circuit would reduce the level of inconsistency in the court's decisions. Analyses of past splits, such as when the Fifth Circuit was split into the Fifth and Eleventh Circuits, could help inform the debate. But regardless of whether one wishes to assess the quality of decision making in an institution or to evaluate an institutional reform, a precondition is an accurate measures of inconsistency.

And yet, accurate measures of inconsistency are rarely available. A version of the fundamental problem of causal inference generally applies in this setting: because we may never observe different administrators or judges independently working on the same case, estimating their disagreement on cases is difficult. Researchers have sought to overcome this difficulty in two main ways. First, they have surveyed judges with simulated case materials, allowing for observation of decisions on the same case (*e.g.*, Dhami 2005). While these inter-rater reliability studies are in high in internal validity, the use of simulated materials can poses serious problems for external validity. For example, if subjects know that they are working on simulated cases, or if they know that their decisions are being scrutinized by researchers, estimates of inconsistency derived from these methods may differ from inconsistency on real-life cases. Indeed, one may worry that such techniques will understate the degree of inconsistency in real-life settings since experimental subjects will often try to be "on their best behavior" and will generally not face the same kinds of caseload burdens that can affect real-world decision making. Moreover, studies like this are expensive, and it is difficult to assure that the simulated materials are representative of the full set of cases.

Second, analysts may try to estimate consistency using real-world cases and decisions. In order to overcome the fact that decision-makers rarely decide the same case, analysts leverage the fact that–due to random or quasi-random assignment of cases to decision-makers–the decision-makers do see the same cases on *average*. This permits an analyst to estimate disparities between decision-makers.(*e.g.*, Ramji-Nogales, Schoenholtz, and Schrag 2007; Nakosteen and Zimmer 2014). For example, if one judge decides (randomly assigned) cases in favor of plaintiffs 30% of the time whereas another judge decides (randomly assigned) cases in favor of plaintiffs 20% of the time, then an analyst could estimate a "disparity" of 10% between these two decision makers. Because these disparity studies rely on real decisions, they are high in external validity. But as we show in more detail below, their reliance on simple differences in means poses serious problems for the internal validity of an inconsistency estimate. For example, two asylum officers may each grant 50% of randomly assigned cases, but it is possible they could disagree as to the correct outcome in 100% of cases. While researchers could manually search for dimensions of disagreement (*e.g.*, perhaps one officer more frequently grants asylum to female applicants, the other officer to male applicants), the important dimensions can be difficult to intuit or detect and can vary between pairs of decision-makers. While such efforts are often an good starting point for estimating inconsistency, we provide a technique for substantially improving upon them.

In this article, we make two main contributions. First, we show how the downward bias in disparity statistics can theoretically be eliminated by targeting a specific kind of heterogeneous treatment effects. Our innovation is to demonstrate that the estimand for inconsistency (formally, the average absolute treatment effect) is equivalent to estimating absolute disparities (formally, absolute average treatment effects) on two "well chosen" subsets of data. In particular, one need only subset based on the direction of disagreement between decision makers. For example, the set of cases where decision maker A is more lenient than decision maker B, and the set of cases where the opposite is true. Second, we demonstrate how machine learning can be used to optimally choose these subsets from observational data. The challenge is that knowing the direction of disagreement on a decision requires knowing the counterfactual decisions made decision makers who were not assigned to a case. This latter step is feasible due to advances in machine learning and increased data availability, which allows us to generate high quality predictions for these unobserved counterfactual quantities. Our method also solves a set of subsidiary technical problems that have plagued studies of inconsistency, including limitations imposed by small sample sizes (*e.g.*, finite sample bias) and imperfect random assignment of cases to decision makers.

We demonstrate our method by estimating inconsistency in the decision-making among the judicial panels of the Ninth Circuit Court of Appeals. To do so, we use an original dataset of all civil appeals heard between 1995 and 2013. To our knowledge, we provide the first robust estimate of inconsistency in the Ninth Circuit, a quantity that is directly relevant to policy debates over the quality of decision making on the court. We find that (1) at least 9% of appeals would be decided differently had they been randomly reassigned and (2) the two most dissimilar kinds of panels decide at least 40% of appeals differently. Moreover, we also show that simple disparity statistics mask at least *half* of this inconsistency. Relying on disparities would therefore lead an analyst or a policy maker to make an overly rosy assessment of the nature of the problem in the Ninth Circuit.

The debate over inconsistency in the Ninth Circuit is ongoing, and the potential policy intervention is still just a hypothetical. However, we provide a crucial component for evaluating such policies: a well-measured outcome variable. Our method for estimating inconsistency has an especially desirable property when evaluating the effects of a policy intervention on inconsistency. By construction, the method is agnostic about the source of inconsistency in an institution, so pre-intervention and post-intervention estimates will not be sensitive to the kind of specification assumptions typically embedded in disparity statistics. For example, suppose that Congress splits the Ninth Circuit and wishes to assess whether inconsistency declined. If one used a disparity statistic to estimate an effect of the policy change—such as the difference in reversal rates before and after the split—then one would need to assume that the main source of disagreement among judges is the same before and after the split. If this is not the case, then a disparity statistic will contain different levels of bias before and after the split. Our statistic, on the other hand, will factor these pre- and post- differences into the estimation and requires only that there are no new *unmeasured* sources of disagreement post-intervention.

2 Quality and Inconsistency

At least since the dawn of the field of public administration, scholars have been concerned with the quality of governance (*e.g.*, Wilson 1887; Taylor 1911). While scholars and wider public often focus their attention on policy making at the highest levels of government, the vast majority of government is devoted to *implementing* policy. This often takes a particular form: "administrators" or "judges" make determinations on a set of "cases." Is company X liable for damages in a tort suit? Is person Y entitled to a social welfare benefit? Does restaurant Z comply with local health ordinances?

To fix ideas, consider a variant of the *case space* used in models of judicial politics (Lax 2011). Each point in the space represents a possible constellation of facts—a case. A decision maker uses a rule, $\hat{x}$ to make one of two possible decisions, which we call "grant" and "deny" (*e.g.*, grant or deny an asylum application). To focus on the core substantive insights, we assume that rules are monotonic in the sense that all cases to the right receive one decision and all cases to the left receive a different decision.[1] The discussion extends outside the context of a unidimensional policy space and monotonic rules, although at the cost of additional complexity that is not relevant for our core argument. Our empirical approach does not invoke these assumptions, even though they are useful for helping us clearly elucidate the core problem our approach solves. In Figure 1, we illustrate this generic

[1] This framework is consistent with a model of decision-making where an official, such as a judge, has single-peaked preferences over rules, but where any individual rule maps cases into a binary decision. See for example, Cameron and Kornhauser (2010).

policy making environment. For simplicity, suppose there are two administrators, 1 and 2.

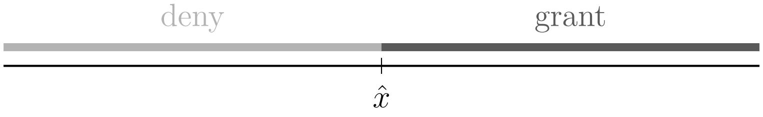


Figure 1: Case Space

Those two administrators may—for whatever reason—use different criteria whenever they make their determinations. For some number of cases, the two administrators disagree, and their decision making is thus inconsistent. A standard way to study this phenomenon using real-life data is through disparity studies. Consider Figure 2. Administrator 1 has a grant rate of around 66% whereas Administrator 2 has a grant rate around 33%. When examining Figure 2, it is apparent that the two administrators disagree on $66 - 33 = 33\%$ of cases. This latter quantity is known as a *disparity.* Many scholars have used disparites to study differences between decision makers across a wide array of institutions, including asylum applications (Ramji-Nogales, Schoenholtz, and Schrag 2007), social security disability appeals (Nakosteen and Zimmer 2014), and judge sentencing (Anderson, Kling, and Stith 1999). The main advantage of disparity studies is that they allow us to study decentralized adjudication systems in as they exist in the real world. But while disparity studies provide us with important information about variation among administrators, they understate how much disagreement there is among those administrators. This is because they rest on the unverified—and often incorrect—assumption that disagreement between decision makers goes in the same direction across all cases. Roughly speaking, this means that two decision makers make their decisions in a "similar" manner, even if they sometimes disagree

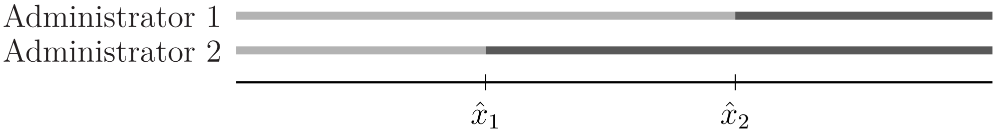


Figure 2: When Disparities Reveal All Inconsistency

on marginal cases. To make this more concrete, suppose that the two administrators do not make their decisions in a similar manner. As depicted in Figure 3, Administrator 1 grants for *high* values in X while Administrator 2 grants for *low* values in X. The two administrators' grant rates remain unchanged, and so does the disparity statistic. But as Figure 3 illustrates, they disagree on many more than 33% of cases. Building on the work of Fischman (2014), we show below that when scholars use disparity as a proxy for disagreement, they are always understating the degree to which the decision makers disagree. Our method minimizes that understatement.

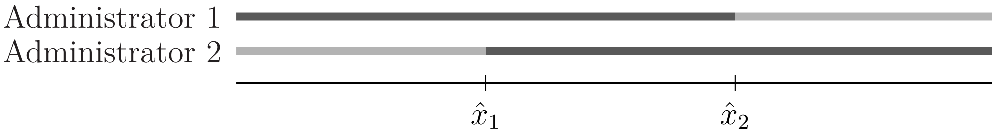


Figure 3: When Disparities Mask Much Inconsistency

3 Measuring Inconsistency

A decentralized adjudication system can be defined by four components. First, there is a finite set of decision makers, which we label $\mathcal{J} = \{1, ..., J\}$, and index by j. Second, there is a finite set of cases about which decisions are made, which we label $\mathcal{N} = \{1, ..., N\}$, and index by i. For a subset of cases decided by a specific decision maker $j \in \mathcal{J}$, we write $\mathcal{N}_j = \{1, ..., N_j\}$. Third, there is a set of possible decisions that could be made for each case, which we label Y.[2] For example, Y could be dichotomous, such as admit/deny or reverse/affirm, or continuous, such as the length of a criminal sentence or a fine.[3] As long as a single determination is made on each case, we allow decision makers to be groups, setting aside the micro-foundations of group decision making. Finally, there is a decision making

[2] We could allow the set of decisions to depend explicitly on the case, but here we assume that the set of possible decisions that could be taken for each case is constant.

[3] We require that the elements of Y be ordinal. Our method does not allow for non-binary categorical outcomes.

function, which is a function mapping sets of decision makers and cases into outcomes, which we denote as $\mathcal{C} : \mathcal{J} \times \mathcal{N} \to Y$. For example, the decision making function for a set of border agents would specify how each agent would decide whether to admit each immigrant they process. Since we only observe actual decisions, we treat $\mathcal{C}$ as a black box and focus on measuring observable patterns in decision making.

Given these components, we can formally define inconsistency in a decision making body. We first must differentiate between two distinct concepts: disagreement and inconsistency. We define *disagreement* to be the proportion of cases where two decision makers would come to different decisions:

$$\delta(j,k) = \frac{1}{N} \sum_{i \in \mathcal{N}} \Big[\mathbf{1}_{(D,\infty)} \big[d(Y_i(j), Y_i(k)) \big] \Big] \tag{1}$$

where $\mathbf{1}$ is the indicator function, $j, k \in \mathcal{J}$ are two different decision makers, $d(\cdot)$ is a metric on Y, and D is a scalar.[4] Since we implement our method using data from the Ninth Circuit, where we treat decisions as binary (*i.e.*, affirm or reverse the lower court's decision), we will assume that $Y = \{0, 1\}$ and we will use the usual Euclidean metric on $\mathbb{R}$. We can thus rewrite equation (1) as:

$$\delta(j,k) = \frac{1}{N} \sum_{i \in \mathcal{N}} \Big[|Y_i(j) - Y_i(k)| \Big] \tag{2}$$

In a decision making body with two decision makers, $\delta(\cdot)$ would completely characterize the amount of inconsistency that exists among the decision makers. However, with more than two decision makers, we must define a composite measure based on the disagreement between each pair. There are many ways to characterize this quantity, but following Fischman (2014), we focus on two: average inconsistency and extreme inconsistency. Define $\mathcal{P}$ to be the set

[4] For example, suppose $Y = [0, 100]$. Then, we might consider two decisions different from one another if they are more than ten units apart. Formally, $d(Y_i(j), Y_i(k)) = |Y_i(j) - Y_i(k)| \leq 10 = D$.

of decision maker pairs: $\mathcal{P} \equiv \mathcal{J} \times \mathcal{J}$. Then, *average inconsistency* is defined by

$$\Delta_a = \frac{1}{|\mathcal{P}|} \sum_{(j,k) \in \mathcal{P}} \delta(j,k) \tag{3}$$

and represents the average level of disagreement between the decision makers. Intuitively, how many decisions would be made differently if all the cases were randomly reassigned? We characterize average inconsistency differently than Fischman (2014). Briefly, we allow that cases could be reassigned to *the same* decision maker, whereas Fischman (2014) assumes that cases are reassigned to different decision makers.

Extreme inconsistency is formally defined by

$$\Delta_e = \max\{\delta(j,k) : (j,k) \in \mathcal{P}\} \tag{4}$$

and represents the disagreement between the two decision makers who are the most dissimilar in their decision making. This quantity can be viewed as bookend normative benchmark, as it captures how high disagreement *could* be.

Limits of Disparities

We have already noted the limitations of disparity studies, but here we show formally how they understate disagreement and inconsistency. As is apparent from equation (1) we express disagreement and inconsistency in idealized terms. In particular, disagreement between two decision makers, j and k, is measured across all cases, even though only only one decision maker can sit on a particular case i. To put this in terms of the Neyman-Rubin causal model, $Y_i(j)$ and $Y_i(k)$ are potential outcomes, where one decision maker is considered the “control” condition and one decision maker is considered the “treatment” condition. As long as the assignment of decision makers is as-if random then standard methods allow us to generate an unbiased estimate of the average treatment effect (ATE, which we denote as $\phi(j,k)$):

$$\phi(j,k) = \mathrm{E}\big[Y(j) - Y(k)\big]$$

In fact, the major *methodological* benefit of studying ATEs is that they are relatively easy to estimate once we satisfy these few assumptions. Specifically, due to the linearity of the estimand, we can decompose it into $\mathrm{E}[Y(j)] - \mathrm{E}[Y(k)]$, which allows us to simply compare the means of the treatment and control groups.

When comparing the decisions of decision makers, researchers usually use this analytical framework to estimate an ATE (or some related quantity, such as regression coefficients).[5] For example, how many more pro-civil rights decisions does an all Democratic panel of judges make as compared to an all Republican panel of judges? Or, how many more asylum applications does Asylum Officer 1 grant than Asylum Officer 2? Such research questions are at least implicitly concerned with measuring the extent of disagreement between decision makers or types of decision makers. However, ATEs can systematically understate actual disagreement, as well as the extent of inconsistency in a decision making system.

Consider an example comparing two state court appellate judges. Suppose one judge is a Democratic and the other is a Republican. Moreover, suppose an analyst is interested in studying how this Democratic judge (D) and this Republican judge (R) decide civil rights cases differently. If she were to try to measure the extent of disagreement between these two judges, the appropriate estimand would be derived directly from equation (2). We label the estimand for equation (2) by $\delta(j,k)$, and in this example it is:

$$\delta(D,R) = \mathrm{E}\big[|Y(D) - Y(R)|\big]$$

[5] There are more examples than we can reasonably list here, but several recent ones are particularly noteworthy. See, for example, Revesz (1997), Anderson, Kling, and Stith (1999), Cockburn, Kortum, and Stern (2003), Farhang and Wawro (2004), Ramji-Nogales, Schoenholtz, and Schrag (2007), Boyd, Epstein, and Martin (2010), and Kastellec (2013).

Of course, estimation of this quantity is complicated by the fact that the expectation operator cannot be linearly decomposed. If instead, the analyst estimates an ATE—which is easier to estimate—then her estimand is

$$\phi(D, R) = \mathrm{E}[Y(D)] - \mathrm{E}[Y(R)]$$

Unfortunately, $\phi(D, R)$ would be a downward biased estimand for disagreement. In Proposition 1 in Appendix E of the supplemental materials, we show generally that[6]

$$\phi(j, k) \leq \delta(j, k). \tag{5}$$

Moreover, equation (5) holds strictly whenever there are "strong heterogeneous treatment effects" as defined by Definition 1 in Appendix E of the supplemental materials. Intuitively, an ATE will always understate disagreement if the treatment has a positive effect in some cases and a negative effect in others. In our example, if Y is whether a plaintiff is victorious in a civil rights case, then an ATE comparing D and R would understate the level of disagreement between them if D is more likely to reverse than R when the defendant won in the lower court but less likely to reverse than R when the plaintiff won in the lower court.

This problem is not solved by re-coding the outcome variable. For example, an analyst might re-code the outcome variable to be pro-plaintiff/pro-defendant instead of reverse/affirm. While this generates a valid measure of the difference in rate of pro-plaintiff decisions for the two judges, it still does not capture disagreement. Suppose, for example, that the two judges differ in their propensity to reverse lower court decisions: D always reverses the lower court decision, and R never does. Moreover, suppose cases won by the plaintiff are appealed as often as cases won by the defendant. Then, the rate of pro-plaintiff

[6] Fischman (2014) demonstrates how measures of inconsistency are always interval-defined, so Proposition 1 can be seen as alternative expression of results from that paper.

decision making by both types of panel is 0.5. An analyst would observe an ATE of zero, potentially concluding that the two judges disagree very little. In fact, they perfectly disagree: *in every case, the panels rule differently.*

As a general principle, estimating ATEs using different outcome variables will reveal different amounts of disagreement between decision makers. The reason is fairly straightforward: each outcome variable reflects disagreement on different dimensions of the decision makers' utility functions. If, for example, preferences about deferring to lower courts is the primary dimension on which appellate judges disagree, then reversal rates will be a better measure of disagreement than whether the plaintiff or defendant ultimately prevail. But, analysts almost never know *ex ante* which outcome best captures disagreement. The estimation of a specific ATE represents a small, and specific, bite of the apple, and may even lead researchers to draw faulty theoretical conclusions. Yet, some ATEs may do a better job of capturing disagreement. For example, the treatment could partition the decision makers into groups that are "most like-minded" and the outcome of interest could be the issue on which the groups of judges disagree most. But, since decision making differs across many possible dimensions, an ATE based on a particular treatment and particular outcome derived intuitively will yield a poor proxy for the overall level of disagreement between judges.

In light of this problem, we reformulate the analysis of decision making as a prediction problem with the goal of backing out the dimensions characterizing the most disagreement.

Getting Around The Problem: Heterogeneous Treatment Effects

One way to view the problem we identify is that there are heterogeneous treatment effects (HTEs, see Athey and Imbens 2015; Grimmer, Messing, and Westwood 2016; Bullock, Green, and Ha 2010). In the context of medicine, for example, a doctor who knows that a particular drug has a positive average treatment effect, may also wish to know which patients respond positively, which respond negatively, and which do not respond at all. Such information, which is thrown away by the particular way that treatment effects are aggregated, has important clinical applications and can help doctors understand better how a drug works. Define $\mathcal{M}$ to be a partition of the set of cases $\mathcal{N}$ that represents a partition of the case-level covariate space and where each $M \in \mathcal{M}$ is nonempty. Then, the estimand of interest is the conditional average treatment effect (CATE):[7]

$$\phi(j, k, M) = \mathrm{E}_M[Y(j) - Y(k)] \tag{6}$$

As Grimmer, Messing, and Westwood (2016) point out, if $\phi(j, k, M)$ varies as M does, then there are heterogeneous treatment effects. In our context, such variation is informative because it allows us to observe how often treatment effects are non-zero, which maps into disagreement. Disagreement for a specific partition $\mathcal{M}$ is:

$$\begin{aligned} \delta(j, k, \mathcal{M}) &\equiv \mathrm{E}_{\mathcal{M}}\big[|\mathrm{E}_M[Y(j) - Y(k)]|\big] \\ &= \mathrm{E}_{\mathcal{M}}\big[|\phi(j, k, M)|\big] \end{aligned}$$

This approach helps researchers avoid making problematic assumptions on the joint distribution of the potential outcomes by bringing the absolute value outside the expectation operator, thus allowing for “traditional” estimation of average treatment effects. The downside to this procedure is that it is highly dependant on the method used to partition the parameter space (*i.e.*, the choice of $\mathcal{M}$). At the limit, $\delta(j, k, \mathcal{M})$ becomes $\delta(j, k)$ as the partition becomes fine enough such that the average treatment effect is estimated for each unit separately (*i.e.*, as $\mathcal{M}$ approaches $\mathcal{N}$). Of course, the fundamental problem of causal inference rules out this possibility (Holland 1986), but one could implement matching to

[7] Another equivalent way to write the CATE is

$$\phi(Y, T, x) = \mathrm{E}[Y(T = 1) - Y(T = 0)|X = x]$$

where X is a vector of covariates and x is a particular value of covariates.

find the closest match for every treated (or control) unit. With a large enough sample size, one could find suitable matches, but the estimation of the average treatment effect for each matched pair would introduce unmanageable finite sample bias.

For practical reasons, an analyst must choose a partition $\mathcal{M}$. Proposition 2 in Appendix E of the supplemental materials shows that all possible partitions $\mathcal{M}$ generate estimands of disagreement that are weakly smaller than the actual level of disagreement. Thus, our task is to pick $\mathcal{M}$ to maximize $\delta(j,k,\mathcal{M})$. We can optimally partition the parameter space into a partition $\mathcal{M}^*$ with exactly two subsets:

$$M^+ = \{i \in \mathcal{N} : Y_i(j) \geq Y_i(k)\} \qquad\qquad M^- = \{i \in \mathcal{N} : Y_i(j) < Y_i(k)\}$$

Then, our estimand for disagreement can be written as

$$\begin{aligned}\delta(j,k,\mathcal{M}^*) &= \mathrm{E}_{\mathcal{M}^*}\left[|\phi(Y,T,M)|\right] \\ &= \mathrm{E}\left[Y(j) - Y(k)|Y(j) \geq Y(k)\right]\Pr[Y(j) \geq Y(k)] \\ &\quad + \mathrm{E}\left[Y(k) - Y(j)|Y(j) < Y(k)\right]\Pr[Y(j) < Y(k)]\end{aligned} \tag{7}$$

Finally, given that our ultimate goal is to estimate inconsistency in an entire decision making body and disagreement only measures inconsistency between two decision makers, the appropriate estimands for average and extreme inconsistency can be written as follows:

$$\Delta_a = \mathrm{E}_{(j,k)\in\mathcal{P}}\left[\delta(j,k,\mathcal{M}^*)\right] \qquad\qquad \Delta_e = \max\{\delta(j,k,\mathcal{M}^*) : (j,k) \in \mathcal{P}\} \tag{8}$$

Estimation

We have shown that ATE-based estimands of disagreement, such as disparities, will always be biased downward. As a result, an ATE-based estimand constitutes a *lower bound* on the true level of disagreement among decision makers. It is important to emphasize once again that attempts to measure disagreement or inconsistency are always either lower or upper bounds on the true measure, including the method we introduce in this paper (see Fischman 2014). However, while our method also yields a lower bound, our contribution is to substantially reduce the bias of more common disparity-type measures of inconsistency. As is apparent from the foregoing discussion, we can only reduce bias by subsetting the parameter space, thus reducing sample sizes and increasing variance. Our estimation challenge is therefore to find the optimal bias-variance trade-off. We face three specific problems in estimation, what we refer to as the problems of *partitioning*, *clustering* and *finite sample bias*.

The partitioning problem refers to the challenge of optimally selecting $\mathcal{M}$ to reduce bias in the estimates for $\delta(j,k)$. The problem is both theoretical and practical. In the previous section, we derive an estimand with the most efficient partition $\delta(j,k,\mathcal{M}^*)$, see equation (7). Because we need only divide our sample into observations where $Y(j) \geq Y(k)$ and $Y(j) < Y(k)$, our partition is as coarse as possible thus increasing variance by as little as possible (relative to the baseline ATE). The practical problem is how to classify observations into the two sets of the partition. We treat this as a prediction problem and recommend using machine-learning methods to generate estimates of $Y_i(j)$ and $Y_i(k)$ for all i that were decided by either j or k. We label these $\widehat{Y}_i(j)$ and $\widehat{Y}_i(k)$. In our illustrative example, described in the next section, we use `Super Learner`, an ensemble method that uses a set of constituent algorithms to predict outcomes in the data (Laan, Polley, and Hubbard 2007).

Until now, our discussion has focused heavily on the estimation of disagreement, $\delta(j,k)$, but not inconsistency. To measure inconsistency, we need to define the set $\mathcal{P}$, which is the set of pairwise comparisons we wish to study. In the context of experiments, this is known as the choice of the treatment arms. This is what we call the clustering problem. In some contexts, the clustering problem is not actually a problem. For example, suppose we have a decision making body with three decision makers, A, B, and C, who each make decisions on 1,000 cases. If we had data from all 3,000 decisions, we would have sufficient sample size to efficiently estimate disagreement between each pair of decision makers: $\delta(A, B)$, $\delta(A, C)$ and $\delta(B, C)$.

This poses a more serious problem where there are a small number of cases assigned to some of the treatments, as predictions would be extremely noisy and uninformative. Consider, for example, the Ninth Circuit. If we define each decision making unit as a specific three-judge panel, then even ignoring senior and designated judges, with 28 active judges there are $3,276$ possible panels. The population of cases is too thinly split among such a large number of decision-making units to allow for meaningful analysis. It will therefore often be necessary to cluster judges into larger groupings to increase sample sizes used to build prediction models. Essentially, we must sometimes re-define the "treatment" and "control" to be panels of different *types* of judges, rather than specific judges. To be clear, clustering explicitly opts for increased bias in order to decrease variance, and the extent to which an analyst trades off bias for variance is context-dependent and discretionary. But we strongly recommend clustering decision makers by similarities in their decision patterns rather than shared demographic or political characteristics. Indeed, as we illustrate below, researchers can use a training set to build decision-predictive models for each decision maker, and decision makers can then be clustered by similarities in the outputs of those models.

Finally, our approach solves a problem identified by Fischman (2014): finite sample bias artificially inflates estimates of inconsistency. Traditional estimates of disagreement overstate inconsistency because they treat all observed differences among decision-makers as reflecting true differences, ignoring the fact that differences are actually a combination of true differences and statistical noise. Moreover, the problem can become more severe as researchers increase variance by subsetting in the search of more inconsistency. Fischman (2014) describes a method for adjusting inconsistency estimates for finite sample bias. Our approach, rather than correcting for finite sample bias, avoids introducing it in the first place. By using training sets to set our expectations for the direction of inter-judge differences *ex* *ante*, we allow noise to result in negative estimates of disagreement when those expectations are not met, eliminating variance's contribution to bias.

4 Application: Ninth Circuit

In the previous section, we described a general approach for estimating inconsistency in an adjudication system. In this section, we demonstrate how to use machine learning to implement our method using a large and extensively coded original dataset of all civil cases filed in the Ninth Circuit and terminated on the merits over a period of nineteen years.[8]

Because the Ninth Circuit decides cases in panels of judges, each of whom only rarely form a panel together, our chosen application is more challenging than standard applications. The infrequency with which the same three judges form a panel means that there is not sufficient data to build predictive models of each panel. Instead, we must somehow group like panels with like panels. In other contexts, where the decision-making units each decide a substantial number of cases, analysts can skip steps one and two of the procedure outlined below.

Our procedure generates estimates of extreme and average inconsistency that uncover a greater degree of disagreement among judges than traditional approaches can. In particular, we find levels of extreme and average inconsistency in the Ninth Circuit of 40% and 9%, respectively.[9] That is, the two most dissimilar types of panels (which are endogenously determined by our machine-learning method) would decide 40% of cases differently, while two randomly selected panels would decide 9% of cases differently on average. As a benchmark, we can compare these estimates to two other disparity measures that cluster judges by party of appointing President, based on theoretical intuitions that partisanship is the dimension

[8] In Appendix B of the supplemental materials, we describe our data and discuss how it constitutes an improvement on other available datasets.

[9] Measures of extreme inconsistency can be overstated because one would expect some regression to the mean. This concern is minimal in our application because the estimated disagreement steadily increases as the panels become more dissimilar.

that captures the most disagreement among judges. Our method substantially outperforms both a naïve comparison of reversal rates and a comparison of pro-plaintiff decision rates, each of which understate inconsistency by at least half. If we use reversal rates, we obtain estimates that uncover substantially less inconsistency: 12% and 4% for extreme and average inconsistency, respectively. If we use pro-plaintiff decision rates, we obtain estimates that uncover even less inconsistency: 2% and 1% for extreme and average inconsistency, respectively.

We now describe the procedure we used to obtain our machine-learning estimates of inconsistency. The procedure follows six basic steps.[10]

1. Build Training-Set Models of Each Judge.
2. Cluster Judges and Apply Categorization to the Training and Test Sets.
3. Use Training Set to Generate Panel-Specific Predictions for the Test Set.
4. Code Test-Set Decisions for Each Pairwise Panel Comparison.
5. Identification Strategy.
6. Estimate Inconsistency Using the Test Set.

Steps 1 and 2 address the clustering problem, using the training set to cluster judges in accordance with their voting patterns rather than their demographic characteristics. Steps 3 and 4 address the partitioning problem, using estimated differences in panel-type voting patterns to partition our data. We again note that Steps 1 and 2 will be unnecessary in

[10] We note that the six steps could be extended to include a seventh step for measuring uncertainty around the point estimates for inconsistency. We focus on point estimates for a number of reasons. First, we do not improve upon the sub-sampling method as described in Fischman (2014). Second, statistical uncertainty with the large data sets that are most appropriate for our technique will be trivial, as were Fischman's (who calculated a standard error of 0.3% for his estimate of inconsistency). Third, highlighting statistical uncertainty can provide a false sense of precision. The important uncertainty in this context concerns the extent to which point estimates remain biased downward, not the statistical uncertainty around those estimates.

systems where decision-making units each decide a large number of cases. Where this is true, there is no reason to waste data by using a designated training set to generate predictions for the test set. Researchers can instead conserve data via cross-validation, allowing parts of the data to successively serve as temporary test sets.

Step 1: Build Training-Set Models of Each Judge

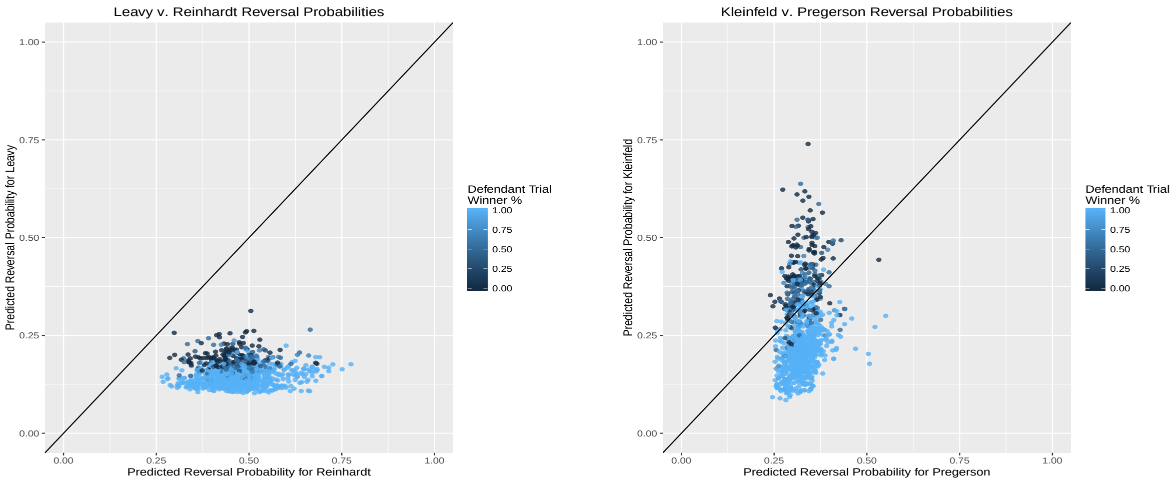
We randomly sample 70% of the data for inclusion in the training set. The remaining 30% is reserved as a test set, which we use to undertake our main analysis. We contribute a greater share of the data to the training set because the tasks we use it for are more data intensive.

Using the training set, we construct a `Super Learner` model of voting for each appellate judge that sat on more than 70 cases. Each judge's model takes as inputs data about each case they sat on, and returns a predicted probability that the judge would vote to reverse the trial court decision. In a sense, the model allows us to characterize each judge's behavior over all their cases leveraging information about the panels they sat on. We include six candidate models in the `Super Learner`: a LASSO regression, the mean reversal rate, two user-specified linear regressions that we thought could capture voting patterns, a random forest, and boosted CARTs. The algorithm then constructs a weighted model of these constituent algorithms that generates the best predictions, as measured by mean squared error. Details are available in Appendix A of the supplemental materials.

It is worth noting that the researcher can be aggressive in this step. If we were presenting predictions from these models as *results* (*i.e.*, as claims about the state of the world), we would have to be concerned that they were an artifact of data mining, intentional or not. But partitioning the test set from the training set allows the researcher to combine the power of clinical judgment and mechanical algorithms—if we get too aggressive and find patterns in the data that reflect chance rather than reality, then applying that "finding" to the test set we set aside will tend to yield uninformative and null estimates. Since we will eventually use these judge-specific models to group judges, if they capture noise rather than signal, then they should not prove useful in the test set.

Figure 4 visualizes judge-specific models for some prominent jurists in the Ninth Circuit. To illustrate the potential benefits of a machine-learning approach, we highlight the relationship between the trial court winner and reversal likelihood. The models suggest that Judge Reinhardt and Judge Leavy decide cases in a highly inconsistent manner but that the magnitude of the overall inconsistency is entirely captured by a comparison of reversal rates—our model predicts that Judge Reinhardt is always more likely to reverse a case than is Judge Leavy. On the other hand, although Judge Pregerson and Judge Kleinfeld have similar overall reversal rates, we predict that they frequently reverse different types of cases. We predict that Pregerson is more likely to reverse when a defendant won in the lower court but less likely to reverse when a plaintiff won.

Figure 4: Comparing Judge Predictions



Step 2: Cluster Judges and Apply to Training and Test Sets

We use each judge's judge-specific model of voting (from Step 1) to generate a predicted probability for how they would have voted in each of the training-set cases, even for cases they did not sit on. Then, for each pair of judges, we calculate the mean absolute distance between the two judges' predicted votes. This is roughly interpretable as an estimate (albeit a very noisy estimate) of the percentage of cases for which that pair of judges would cast different votes. Using these pairwise distances between judges, we use standard cluster analysis to group judges.[11] The judges cluster pretty clearly into six groups. With more data, it might make sense to use a finer clustering, but anything more than six groups begins to stretch the data too thinly, leaving too few observations of each panel type to make statistically relevant comparisons.

We take the liberty of using the names of well-known judges to label the clusters: judges are thus each identified as being part of the Reinhardt Cluster, the Leavy Cluster, the Kozinski Cluster, the Pregerson Cluster or the O'Scannlain Cluster. The exact membership and demographic characteristics of the groups are available in Tables 3 to 8 in Appendix C of the supplemental materials. We also add a cluster that we label the Visiting Cluster. This cluster consists of judges who had fewer than 70 observations in the training set, most of whom are judges sitting by designation. Throughout the text, we refer to particular "types" of judges using formatted labels corresponding to the cluster names: `R`, `L`, `K`, `P`, `O` and `V`. For example, we refer to a judge from the O'Scannlain Cluster as an `O`-judge, and a panel of such judges as an `OOO`-panel.

Step 3: Generate Panel-Specific Predictions and Apply to Test Set

We reuse the training set and the same candidate models that we used to build the judge-specific models (Step 1) to generate *panel-specific* predictions for each case in the test set.[12]

[11] Specifically, we use the `hclust` package for hierarchical clustering in `R`. We use the `ward.D` method for its tendency to generate clusters of relatively equal size.

[12] Ideally, one would use a new training set to construct panel-specific predictions, as any noise that contributed to the grouping of judges could be compounded in the panel model, leading us to over-estimate differences between panel types. But we think it was more important to "spend" our data on the judge-specific models. Furthermore, any over-estimating of differences between panels will ultimately bias our test-set estimates of inconsistency downward.

As we show in Table 1, our `Super Learner` model outperforms each constituent algorithm.

Table 1: Model Performance

Model	MSE	Weight
Super Learner	0.179	–
Boosted Trees	0.181	0.51
Random Forest	0.181	0.45
LASSO	0.184	0.00
Regression 1	0.189	0.00
Regression 2	0.189	0.00
Regression 3	0.185	0.04
Mean	0.195	0.00

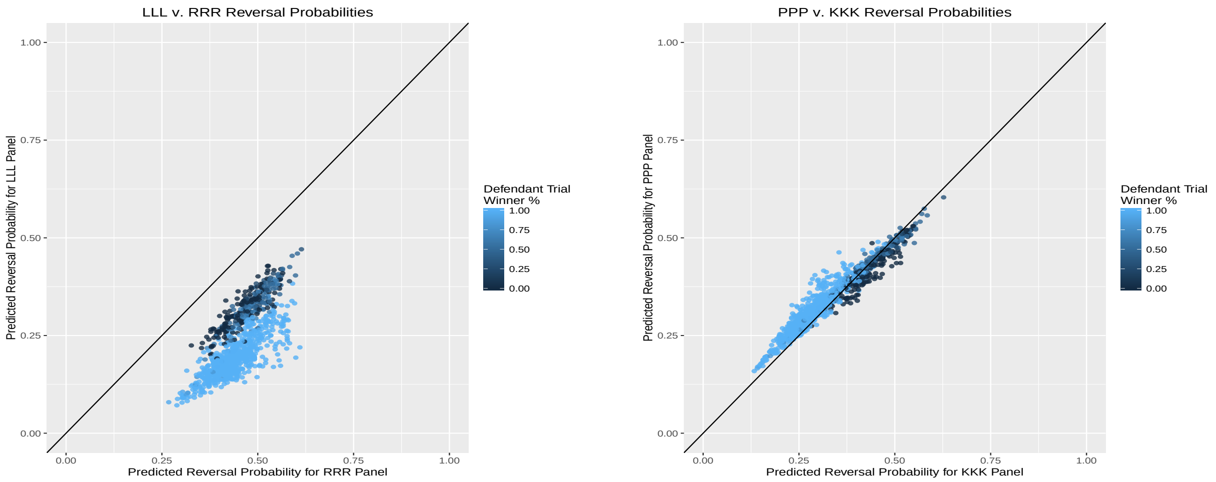
Figure 5 visualizes predictions for four of the panel types and highlights the potential for machine-learning approaches. The left panel suggests that the main source of disagreement between `LLL`-panels and `RRR`-panels is something related to judges' reversal proclivity. Even though `LLL`-panels appear to exhibit a pro-defendant leaning, `RRR`-panels do not. `RRR`-panels simply reverse *a lot* more than `LLL`-panels, regardless of whether the plaintiff or defendant won in the lower court. An analyst concerned about the mechanisms driving differences between `LLL`-panels and `RRR`-panels would infer that willingness to reverse is what actually differentiates decision making between `RRR`- and `LLL`-panels.

On the other hand, the right panel of the figure illustrates that substituting a `KKK`-panel for a `PPP`-panel is predicted to decrease the probability of reversal where a defendant won in lower court but increase the probability where a defendant lost. The figure suggest that a simple comparison of average reversal rates of `PPP`-panels and `KKK`-panels would understate the extent to which the two types of panels decide cases differently because they tend to reverse *different types of cases.* In other words, judges who are more similar to Kozinski tend to reverse defendant-wins less often than judges more similar to Pregerson, and vice versa.

Moreover, if Steps 1 and 2 are not necessary (as explained above), the panel-specific predictions can be estimated on the entire dataset.

Next, in Step 4, we show how these predictions can be used to re-code decisions in the test set so that we can more accurately estimate the actual level of inconsistency in adjudication.

Figure 5: Comparing Panel Predictions



Step 4: Code Test Set Outcomes for Each Pairwise Panel Comparison

As we discussed in Section 3, the presence of heterogeneous treatment effects means that a comparison of overall reversal rates between types of panels will lead one to underestimate the level of inconsistency in adjudication. For example, all-Republican panels and all-Democratic panels might have identical reversal rates, but all-Republican panels may be more likely to reverse civil rights cases when plaintiffs won in the lower court and all-Democratic panels may be more likely to reverse civil rights cases when defendants won. As a result, our challenge is to optimally partition the parameter space to capture the most amount of disagreement.

We seek to estimate $\delta(j, k, \mathcal{M}^*)$ for each pair of decision makers, $(j, k) \in \mathcal{P}$. Restating equation (7), our estimand is:

$$\delta(j,k,\mathcal{M}^*) = \mathrm{E}\left[Y(j) - Y(k) | Y(j) \geq Y(k)\right] \Pr[Y(j) \geq Y(k)]$$
$$+ \mathrm{E}\left[Y(k) - Y(j) | Y(j) < Y(k)\right] \Pr[Y(j) < Y(k)]$$

We allow `Super Learner` to acquire knowledge about how panel types decide cases. Specifically, we estimate potential outcomes on our training set, $\widehat{Y}_i(j)$ for all $i \in \mathcal{N}$ and all $j \in \mathcal{P}$. Then, for a given pairwise comparison (j,k), we code the outcome of a case as a j-decision or a k-decision depending on whether a j panel or k panel is predicted as more likely to have made the decision that was actually made (according to the model that we constructed with the training set). To implement this, we treat either j or k as the "treatment" and code a new outcome variable, which we refer to as "autocoded" and label $\widetilde{Y}_i^{j,k}(\cdot)$. Suppose we define j to be the treatment, then:

$$\widetilde{Y}_i^{j,k}(\cdot) = \begin{cases} Y_i & \text{if } \widehat{Y}_i(j) \geq \widehat{Y}_i(k) \\ 1 - Y_i & \text{if } \widehat{Y}_i(j) < \widehat{Y}_i(k) \end{cases} \tag{9}$$

One can interpret $\widetilde{Y}_i^{j,k}(\cdot)$ to be whether or not observation i had a j-like outcome. For example, if j is an LLL-panel and k is an RRR-panel, $\widetilde{Y}_i^{\texttt{LLL},\texttt{RRR}}(\cdot) = 1$ implies that observation i featured an outcome Y_i that was consistent with a model of LLL decision making, but not RRR decision making. Notice that our autocoded outcome estimates $\delta(j,k,\mathcal{M}^*)$ by splitting our data for each pairwise comparison of panels into four groups:

$$\widehat{M}_j^+ \equiv \{i \in \mathcal{N}_j : \widehat{Y}_i(j) \geq \widehat{Y}_i(k)\} \qquad \widehat{M}_j^- \equiv \{i \in \mathcal{N}_j : \widehat{Y}_i(j) < \widehat{Y}_i(k)\}$$
$$\widehat{M}_k^+ \equiv \{i \in \mathcal{N}_k : \widehat{Y}_i(j) \geq \widehat{Y}_i(k)\} \qquad \widehat{M}_k^- \equiv \{i \in \mathcal{N}_k : \widehat{Y}_i(j) < \widehat{Y}_i(k)\}$$

Our estimator is therefore:[13]

$$\widehat{\delta}(j,k,\mathcal{M}^*) = \frac{1}{N_j}\left(\sum_{\widehat{M}_j^+} Y_i + \sum_{\widehat{M}_j^-}(1-Y_i)\right) - \frac{1}{N_k}\left(\sum_{\widehat{M}_k^+} Y_i + \sum_{\widehat{M}_k^-}(1-Y_i)\right)$$

In Step 6, we go into more detail about how we estimate our composite measures of inconsistency, Δ_a and Δ_e, using this procedure.[14]

Step 5: Identification Strategy

Our analysis relies on an assumption of statistical independence between cases and judge panels. We require this assumption since our goal is to isolate the *unconfounded* effect of judges on outcomes. If this were not the case, our estimates of each judge's effect on case outcomes could simply reflect differences in the types of cases that judges are assigned. However, as we discuss in Appendix D in the supplemental materials, there are at least two possible threats to the randomization assumption in the context of the Ninth Circuit. In fact, proper randomization is rarely guaranteed in decision making systems, so this step provides a technique for correcting for potential selection bias.

In order to guard against threats to randomization, we move beyond raw comparisons of panel decision rates (that would rely on random assignment) and account for the possibility that some panels may be more or less likely to issue decisions in cases that are, as a general matter, more or less likely to be reversed. Because bias occurs when a confounding variable is correlated with both the treatment and the outcome, we use a prognostic score correction, which aims to make the confounding variable orthogonal to the outcome (Hansen 2008).[15]

[13] See Proposition 3 in Appendix E of the supplemental materials, where we prove that our autocode procedure generates an equivalent estimand to $\delta(j,k,\mathcal{M}^*)$.

[14] Recall, the estimate of $\widehat{\delta}(j,k,\mathcal{M}^*)$ is for a pairwise comparison of j and k type panels. We seek an *overall* measure of inconsistency, where we incorporate the $\widehat{\delta}(j,k,\mathcal{M}^*)$ for each pairwise comparison.

[15] Researchers have traditionally used propensity scores (or some other technique) to force independence between the confounding variable and the treatment.

To do this, we use machine learning and the training set to estimate each case's predicted probability of reversal under the "control" condition, which we have been denoting by k. We label the predicted probability $\widehat{\psi}_i(k, \mathbf{X}_i)$, which is commonly referred to as the "prognostic score" (Hansen 2008). In our main analysis, we incorporate these prognosis scores directly into our outcome variable. That is, instead of using the actual decision,

$$Y_i = \begin{cases} 1 & \text{if panel reverses} \\ 0 & \text{if panel affirms,} \end{cases}$$

we use the difference between the predicted probability of a reversal under the control condition (estimated with the training set, and referred to as k) and the actual outcome,

$$Z_i^{j,k} \equiv Y_i - \widehat{\psi}_i(k, \mathbf{X}_i).$$

Since $Z_i^{j,k} \in [-1, 1]$, the bias corrected autocode is

$$\dot{Y}_i^{j,k} = \begin{cases} Z_i^{j,k} & \text{if } \widehat{Y}_i(j) \geq \widehat{Y}_i(k) \\ -Z_i^{j,k} & \text{if } \widehat{Y}_i(j) < \widehat{Y}_i(k) \end{cases}$$

Thus, if, for example, some panels are more likely to issue decisions in cases with high reversal probabilities (due to breakdowns in randomization), that fact is accounted for (as best as possible) when making comparisons.[16] Our modified estimator, now resistant to

[16] Two details are worth mentioning. First, we estimate the prognosis scores with and without party variables for fear they could be post-treatment. The results do not change significantly. Second, since the identification of the "treatment" and "control" groups is arbitrary when comparing two panels, all of our analyses with prognosis scores is completed twice, with each panel being regarded as the "control" group. Results are not sensitive to the arbitrary choice of the "control" group, but we nevertheless average results.

breakdowns in the randomization procedure, is:

$$\widehat{\delta}(j,k,\mathcal{M}^*|\mathbf{X}) = \frac{1}{N_j}\left(\sum_{\widehat{M}_j^+} Z_i^{j,k} - \sum_{\widehat{M}_j^-} Z_i^{j,k}\right) - \frac{1}{N_k}\left(\sum_{\widehat{M}_k^+} Z_i^{j,k} - \sum_{\widehat{M}_k^-} Z_i^{j,k}\right)$$

Step 6: Estimate Inconsistency with the Test Set

Now we estimate average inconsistency and extreme inconsistency. Our estimator for extreme inconsistency is straight forward:

$$\widehat{\Delta}_e \equiv \max\left\{\widehat{\delta}(j,k,\mathcal{M}^*|\mathbf{X}) : (j,k) \in \mathcal{P}\right\} \tag{10}$$

Recall that average inconsistency is an estimate of the percentage of cases that would have been decided differently if the court had re-randomized the assignment of cases to panels. In calculating average inconsistency, we account for the fact that different panel types hear greater or fewer cases using a re-randomization weighting. Consider panels of type j and k. Then, the probability that a j panel is re-randomized a k panel (or vice versa) is:

$$\widehat{w}(j,k) = \frac{N_j}{N} \cdot \frac{N_k}{N}$$

where N_j and N_k are the number of cases seen by a type j panel and type k panel, respectively. Our estimator for average inconsistency is therefore:

$$\widehat{\Delta}_a \equiv \sum_{(j,k)\in\mathcal{P}} \widehat{w}(j,k)\ \widehat{\delta}(j,k,\mathcal{M}^*|\mathbf{X}) \tag{11}$$

With these six steps, a researcher can generate estimates of inconsistency in adjudication systems, and, as we've detailed above, can offer substantial improvements to existing methods.

5 Discussion: Ninth Circuit Inconsistency

According to many of its critics, the Ninth Circuit is too inconsistent in its decision making. A first step in assessing whether there is *too much* inconsistency is getting measurements on *how much* there is. Critics also propose that splitting the court would reduce inconsistency. Again, a first step in empirically evaluating the success of a split will be getting before and after measurements of inconsistency. In this article, we developed a theory and technique for accomplishing these first steps with observational data. Below, we briefly discuss how one might assess whether the level of inconsistency is excessive and whether a split was effective.

Assessment

Scholars have long used measures of inter-judge disparities to assess the quality of adjudication systems. Although they have so far used only unidimensional measures of disagreement that likely understate the extent of inconsistency, some systems—such as those for making decisions in asylum and social security disability awards—are home to such large and easily detectable disparities that most people seem to agree: something is not right. But in many contexts, it will not be immediately clear that there is a problem worth trying to address. It is not obvious to us, for example, that something is amiss with the Ninth Circuit given our findings that panels could disagree on at least 40% of cases and that at least 9% of cases would be decided differently if they were randomly reassigned. Is that level of inconsistency disconcerting?

Answering the question is important. Inconsistency has been at the heart of the debate over whether the Ninth Circuit should be split into two or more smaller circuits. Ninth Circuit Judge Tallman, for example complained about the inconsistency in the Ninth Circuit in Congressional testimony:

> The problem that we have now with 50 judges resident, active and senior, and 150 to 200 visiting judges is that it is like going to Las Vegas in terms of what the outcome is going to be.

But others have vigorously disagreed with Judge Tallman's perspective. The Federal Judicial Center, for example, has claimed that "despite concerns about the proliferation of precedent as the courts of appeals grow, there is currently little evidence that intracircuit inconsistency is a significant problem" (McKenna 1993).

Can measures of inconsistency help to illuminate this debate? We think so, and we propose three potential benchmarks for aid in determining whether estimates of inconsistency are alarming. First, insider intuitions might be useful: empirical evidence of inconsistency might convince judges that inconsistency is more prevalent and serious than they think it is. Second, we can use public claims about the percentage of "hard cases" to gauge whether the Ninth Circuit has a problem. Prominent legal figures like Judge Harry Edwards and Judge Cardozo have claimed that between 5-15% of cases are legally indeterminate. Our evidence shows that some panels can disagree on at least 40% of cases, which suggests a much higher level of indeterminacy in the Ninth Circuit.[17] Finally, we could compare estimates of inconsistency in the Ninth Circuit to the other circuits. Unfortunately, due to restrictions placed on access to the court's public records system (PACER) we do not have data for other circuits. While these comparisons would not be definitive—the differences in estimates might reflect different case composition and/or differences in the detectability of inter-judge disagreement—much higher levels of detectable inconsistency in the Ninth Circuit would be cause for concern.

[17] It is also possible that 40% of cases are not legally indeterminate: some fraction of that disagreement could stem from judicial error. But we find the indeterminacy/error dichotomy generally unhelpful—one judge's error is often another judge's indeterminacy.

Evaluating Institutional Reform

We might want to know whether institutional reforms increase or decrease inconsistency. As Fischman (2014) has pointed out, there are challenges:

> If one goal of an institutional reform is to reduce inconsistency, it will be difficult to assess whether the reform is successful. Estimates of inconsistency pre- and post-intervention will be interval-identified, and these intervals will typically overlap. When this occurs, it will be impossible to determine whether the intervention increased or decreased inconsistency. Thus, evaluation of institutional changes will typically require additional data and assumptions.

Our approach to measuring inconsistency efficiently uses all available data in an effort to make any required additional assumptions as unobjectionable as possible. And we think those additional assumptions will often be reasonable. In brief, in order for a reduction in estimated inconsistency to reflect an actual decrease in inconsistency levels, it must be the case that the reform did not increase inconsistency on new, undetectable dimensions. While it may always be possible that there are substantial deviations from that assumption (*i.e.*, some significant increases in disagreement are undetectable), our data-adaptive approach minimizes the possibility. As a result, our method provides an advantage over using disparity statistics to evaluate institutional reforms. Since disparities naturally build-in assumptions about the source of disagreement between decision makers, institutional reforms that alter the decision making environment in important ways may also shift the dimensions that best characterize disagreement. A disparity statistic does not account for this; our measure of inconsistency does. In other words, our method, unlike disparity statistics is not subject to specification bias.

To make this point more concrete, consider the hypothetical example of splitting the Ninth Circuit. Suppose the circuit is split in such a way that the two new circuits have equal numbers of Democratic and Republican appointees.[18] In this situation, the new partisan

[18] For example, if an incumbent Republican president gets to appoint a bunch of judges to newly created

make-up of the circuits may change the main dimension on which judges in those circuits disagree. In our analysis, we discovered that differences in *reversal rates* explain more disagreement among panels than differences in *pro-plaintiff decision rates.* However, if the partisan balance shifts after a hypothetical circuit split, it is plausible that the empirical pattern we identified reverses itself. Then, if pro-plaintiff decision rates explain more disagreement than reversal rates after the circuit split, the estimated effect of the split using a single disparity statistic pre- and post-split will be biased. For example, if an analyst used a reversal rates disparity statistic, they would overestimate the degree to which the institutional reform actually reduced inconsistency. Assuming such changes are perceptible in the available data, our method will account for them and thus automatically correct for potential biases introduced by changes in the courts' composition.

6 Conclusion

We have presented a method for estimating inconsistency in decentralized adjudication systems. Although it is still only a lower bound on inconsistency, it represents a vast improvement on the disparity studies that have so far been used. Particularly when coupled with inter-rater reliability studies that use simulated case materials, our observation-based method provides a way forward for the rigorous study of inconsistency. We applied our method to the decision making in the Ninth Circuit, showing how we can start building measures of performance.

References

Alameda County Grand Jury. 2011. *2010-2011 Alameda County Grand Jury Final Report.* Technical report. `http://www.acgov.org/grandjury/final2010-2011.pdf`.

judgeships in the two circuits.

Anderson, James M., Jeffrey R. Kling, and Kate Stith. 1999. “Measuring Interjudge Sentencing Disparity: Before and After the Federal Sentencing Guidelines.” *Journal of Law and Economics* 42 (S1): 271–308.

Athey, Susan, and Guido W. Imbens. 2015. “Machine Learning Methods for Estimating Heterogeneous Causal Effects.”

Boyd, Christina L., Lee Epstein, and Andrew D. Martin. 2010. “Untangling the Causal Effects of Sex on Judging.” *American Journal of Political Science* 54 (2): 389–411. doi:`10.1111/j.1540-5907.2010.00437.x`.

Bullock, John G., Donald P. Green, and Shang E. Ha. 2010. “Yes, But What’s the Mechanism? (Don’t Expect an Easy Answer).” *Journal of Personality and Social Psychology* 98 (4): 550–558.

Cameron, Charles M., and Lewis A. Kornhauser. 2010. “Modeling Collegial Courts (3): Adjudication Equilibria.” `http://ssrn.com/abstract=2153785`.

Cockburn, Ian, Samuel Kortum, and Scott Stern. 2003. “Are All Patent Examiners Equal? Examiners, Patent Characteristics, and Litigation Outcomes.” In *Patents in the Knowledge-Based Economy,* 19–53. Washington, DC: The National Academic Press.

Dhami, Mandeep K. 2005. “From Discretion to Disagreement: Explaining Disparities in Judges’ Pretrial Decisions.” *Behavioral Sciences and the Law* 23 (3): 367–386.

Farhang, Sean, and Gregory J. Wawro. 2004. “Institutional Dynamics on the U.S. Court of Appeals: Minority Representation Under Panel Decision Making.” *Journal of Law, Economics, and Organization* 20 (2): 299–330.

Fischman, Joshua B. 2014. “Measuring Inconsistency, Indeterminacy, and Error in Adjudication.” *American Law and Economics Review* 16 (1): 40–85.

Grimmer, Justin, Solomon Messing, and Sean J. Westwood. 2016. “Estimating Heterogeneous Treatment Effects and the Effects of Heterogeneous Treatments with Ensemble Methods.”

Hansen, Ben B. 2008. “The Prognostic Analogue of the Propensity Score.” *Biometrika* 95 (2): 481–488.

Ho, Daniel E. 2017. “Does Peer Review Work? An Experiment of Experimentalism.” *Stanford Law Review* 69:1–119.

Holland, Paul W. 1986. “Statistics and Causal Inference.” *Journal of the American Statistical Association* 81 (396): 945–960. doi:10.2307/2289064.

Kahneman, Daniel, Andrew M. Rosenfield, Linnea Gandhi, and Tom Blaser. 2016. “Noise: How to Overcome the High, Hidden Cost of Inconsistent Decision Making.” *Harvard Business Review* (October): 36–43.

Kastellec, Jonathan P. 2013. “Racial Diversity and Judicial Influence on Appellate Courts.” *American Journal of Political Science* 57 (1): 167–183. doi:10.1111/j.1540-5907.2012.00618.x.

Laan, Mark J. van der, Eric C. Polley, and Alan E. Hubbard. 2007. “Super Learner.” http://biostats.bepress.com/ucbbiostat/paper222/.

Lax, Jeffrey R. 2011. “The New Judicial Politics of Legal Doctrine.” *Annual Review of Political Science* 14:131–157. doi:10.1146/annurev.polisci.042108.134842.

Lipsky, Michael. 1980. *Street-Level Bureaucracy: Dilemmas of the Individual in Public Services.* New York: Russell Sage Foundation.

McKenna, Judith A. 1993. *Structural and Other Alternatives for the Federal Courts of Appeals.* Technical report. Washington, DC: Federal Judicial Center.

Nakosteen, Robert, and Michael Zimmer. 2014. “Approval of Social Security Disability Appeals: Analysis of Judges’ Decisions.” *Applied Economics* 46 (23): 2783–2791.

Narea, Nicole. 2017. *Iranian National Challenges USCIS Investor Visa Denial.* Accessed June 26, 2017.

Ornstein, Charles, and Lena Groeger. 2012. *Two Deaths, Wildly Different Penalties: The Big Disparities in Nursing Home Oversight.*

Ramji-Nogales, Jaya, Andrew I. Schoenholtz, and Phillip G. Schrag. 2007. "Refugee Roulette: Disparities in Asylum Adjudication." *Stanford Law Review* 60:295–412.

Revesz, Richard L. 1997. "Environmental Regulation, Ideology, and the D.C. Circuit." *Virginia Law Review* 83 (8): 1717–1772.

Taylor, Frederick Winslow. 1911. *The Principles of Scientific Management.* New York: Harper & Brothers Publishers.

Trump v. Int'l Refugee Assistance Project. 2017. 582 U. S. ______. LexisNexis Academic (June 27, 2017).

Turnbull, Lornet. 2013. *Suspicious Feds Turn Back Many Foreigners at Airport.* `http://www.seattletimes.com/seattle-news/suspicious-feds-turn-back-many-foreigners-at-airport/`.

Wilson, Woodrow. 1887. "The Study of Administration." *Political Science Quarterly* 2 (2): 197–222.