Mood and the malleability of moral reasoning: the impact of irrelevant factors on judicial decisions

Daniel L. Chen[*], Markus Loecher[†]

July 1, 2023

Emotions are said to underlie moral decision-making. We detect intra-judge variation spanning three decades in 1.5 million judicial decisions driven by factors unrelated to case merits. U.S. immigration judges grant an additional 1.4% of asylum petitions–and U.S. district judges assign 0.6% fewer prison sentences and 5% longer probation sentences—on the day after their city's NFL team won, relative to days after the team lost. Bad weather has the opposite effect of a team win. Unrepresented parties in asylum bear the brunt of NFL effects. The effect on district judges only appears for judges born in the same state as the current state of residence, providing clean evidence of extraneous influences on judge decision-making as opposed to lawyer or applicant behavior.

Moving beyond OLS, we utilize models from machine learning to estimate the sentence length relative to the sentencing guideline. We find that while several appropriate features predict sentence length, such as details of the crime committed, other features seemingly unrelated, including daily temperature, sport game scores, and location of trial, are predictive as well. The predictive power of the unrelated events is derived from the permutation based variable importance score in random forests. We address recent criticism of the reliability of these scores with double residualization.

[*] Daniel L. Chen, daniel.chen@iast.fr, Toulouse School of Economics, Institute for Advanced Study in Toulouse, University of Toulouse Capitole, Toulouse, France. We thank research assistants for contributions to the project and numerous colleagues with helpful comments. This project was conducted while Chen received financial support from the European Research Council (No. 614708), Swiss National Science Foundation (No. 100018-152768), and Agence Nationale de la Recherche.

[†] Berlin School of Economics and Law, Germany

1. Introduction

What determines judicial decisions? We would like to believe it is “the law". To be sure, the law may be hard to determine or even indeterminate. The identity of the judge can thus make a large difference. This simple fact was statistically established at least a century ago (Everson [1919]) and triggered policy responses such as the U.S. Federal Sentencing Guidelines. In particular, judicial decisions differ by judicial ideology, as surveyed in later studies (Fischman and Law [2009]) and recognized in the frequent judicial confirmation battles. But while these inter-judge differences show that the meaning of “the law” is not unique in practice, they are consistent with each judge consistently applying his or her version of the law. Indeed, normative theory readily admits such variation because the correct interpretation of the law is not unique, or at least not discernible for real world judges (Dworkin [1986], Kennedy [1998]).

In this article, we provide evidence from a natural experiment for a different kind of variation. we detect intra-judge variation driven by factors completely unrelated to the merits of the case, or to any case characteristics whatsoever. Concretely, we show that asylum grant rates in U.S. immigration courts differ by the success of the court city’s NFL team on the night before, and by the city’s weather on the day of the decision. Our data include 1.5 million decisions spanning three decades – 22,000 asylum decisions on Mondays after a game; a half million asylum decisions in total; and a million sentencing decisions – and allows exclusion of confounding factors, such as scheduling and seasonal effects. Most importantly, the design holds the identity of the judge constant. On average, U.S. immigration judges grant an additional 1.4% of the asylum petitions–and U.S. district judges assign 0.6% fewer prison sentences and 5% longer probation sentences (a substitute for imprisonment)—on the day after their city’s NFL team won, relative to days after the team lost. Bad weather on the day of the decision has approximately the opposite effect. By way of comparison, the average grant rate is 39%, the average imprisonment rate is 88%, and the average probation length is 40 days. Effects are larger with upset losses (defeats when the team was predicted to win by four or more points), but not upset wins (victories when the team was predicted to lose), consistent with asymmetry in the gain-loss utility function.

Unlike previous studies, the available data allow us to determine if sports outcomes and weather influence the judge directly or indirectly through lawyer behavior—we find that the effect of NFL outcomes on asylum decisions is entirely borne by unrepresented applicants. In this sample, there is no lawyer. Moreover, we suspect that refugees are not avid fans of NFL games to be affected by their outcomes. In U.S. federal district courts, sentencing decisions are almost always made after a guilty plea, without a trial. Furthermore, we have the birth state of the district court judges, and we find that the effect of NFL outcomes on sentencing decisions is present only for those judges born in the same state as the current state of residence, which points towards a more direct effect on the judge rather than an indirect effect. Taken together, our results demonstrate that case outcomes depend on more than “the law,” “the facts” of the case, judicial ideologies, or even constant judicial biases, for example with respect to race. While we suspect that many practitioners would not be surprised by that basic claim, the contribution of this article, however, is to provide clear causal identification of two such factors, and to measure their magnitude. The measured effects of 1.4%, 0.6%, and 5% appear large for two reasons. First, we can only measure one emotional influence out of many.

Presumably, other factors such as family problems or joys, traffic jams, or health fluctuations have an even greater influence on a judge's state of mind and thus plausibly case outcomes. If we had data on these, we would expect to find much larger effects. Second, even the estimate of NFL effects is only a lower bound, for lack of better data on the diverging preferences of the judges. This introduces measurement error and biases the coefficient towards zero.

Numerous field studies that have shown humans in many settings to be influenced by seemingly irrelevant factors in general and by sports outcomes and weather in particular. For example, sports results affect stock returns (Edmans et al. [2007]), sports results influence voting in political elections (Healy et al. [2010]), and disappointing NFL football results trigger domestic violence (Card and Dahl [2011]). Many similar studies exist for weather, and the research on weather's effect on economics and finance has been summarized and experimentally traced to weather's effects on risk attitudes (Bassi et al. [2013]). Moreover, bad weather on visiting days increases the chance that an admitted student will enroll (Simonsohn [2010]). Such effects are manifestations of the broader point that weather strongly influences mood (Connolly [2013]).

One article detects time-of-day patterns in Israeli judges' parole decisions (Danziger et al. [2011b]). Concretely, the article shows that parole approval rates drop with the time from the judges' last meal. One potential problem with this research design is that the order of prisoners' appearance before the judges and the exact time judges choose to take breaks may not be random (Weinshall-Margel and Shapard [2011], Danziger et al. [2011a]). The sample size ($N = 1,112$ and 8 judges) is also several orders of magnitude smaller than ours. In parallel work, a second article finds that outcomes of games played by Louisiana State University football team affects judicial decisions handed down by judges in a Louisiana juvenile court (Eren and Mocan [2016]). The article finds that unexpected losses increase sentence length on juvenile defendants imposed by the judges by around 6.4 percent. Our complementary analysis yields similar results using the same regression specification for upset losses. Like their analysis finding larger effects for judges who attended Louisiana State University, we find larger effects for judges who grew up in the area. One difference between their setting and ours is their sample size of 9,346 defendants and 207 judges is smaller than the 1.5 million decisions and 1,684 judges analyzed in this article. A second difference is that judges in Louisiana juvenile courts may be somewhat less professional than judges appointed by the U.S. President and confirmed by the Senate, which may amplify emotional influences.[1] A third article, also in parallel, finds that asylum denial rates monotonically increase with temperature. (Heyes and Saberian [2018])[2] Our complementary analysis finds temperature

[1] Another difference is opposing interpretations on probation length, which is authorized by U.S. law as an alternative for imprisonment and viewed as an act of grace, delaying the imposition or execution of a sentence. Eren and Mocan [2016] interprets probation as a measure of severity. At least in our setting of the federal courts, it is not so clear. In general, probation can be interpreted as the judge viewing the criminal record of the defendant as not sufficient for imprisonment of a certain length, or as a form of rehabilitation. Historically, a defendant could be assigned a sentence and be placed on probation, with his sentence suspended. In *Davis v. Parker*, 293 F Supp 1388 (DC Del 1968), probation was "an act of grace". In *United States v. Allen*, 349 F Supp 749 (ND Cal 1972), the court ruled that "Probation's primary objective is to protect society by rehabilitating the offender".

[2] Heyes and Saberian [2018] use data from "a website run by an international consortium of agencies that helps asylum seekers in Australia, Canada, the United States and several countries in Europe." This data source

effects for sentencing decisions. Their article attributes the channel to the judge (rather than the lawyer or defendant) because of the differences by gender of judge. A potential alternative explanation could be that male judges are less affected by the different behaviors that lawyers and applicants exhibit on hot days. High temperature increases apathy and lowers effort (Cao and Wei [2005], Wyndham [2013]). We also complement their evidence with a less ambiguous measure of bad weather (rain, winds, snow), whereas one might expect non-monotonic mood effects with temperature (which we find with a larger sample). Baylis [2018] documents a clear U-shape between temperature and sentiment measured in twitter. None of these three articles utilize models from machine learning.

There are also various papers showing clear judicial biases in the laboratory environment (e.g., Guthrie et al. [2000], Guthrie et al. [2007]; Rachlinski et al. [2009], Rachlinski et al. [2013]; cf. Simon [2012]). In particular, these experiments clearly identify racial bias (e.g., Rachlinski et al. [2009]). Outside the lab, findings of racial bias are always subject to at least the theoretical possibility that different outcomes reflect unobserved case heterogeneity beyond race.

Massive inter-judge variation in asylum grants has been documented by Ramji-Nogales et al. [2007], who introduced the legal literature to the asylum data. They showed that grant rates for the same applicant nationality in the same city could be anywhere between, e.g., 0 and 68% depending on the judge who heard the case. Our findings complement theirs. Their findings, while shocking, would be consistent with individual judges steadily applying the same legal philosophy – their own –, but legal philosophy differing across judges. By contrast, our finding shows that consistency is limited even within judge.

Asylum courts involve serious, potentially life-or-death decisions (Ramji-Nogales et al. [2007]). Their case load is very high, forcing immigration judges to make important decisions with little time (on average 7 minutes by one estimate[3]) and hence presumably with less deliberation and more of a "hunch" than other judges (Hutcheson, Jr. [1929]). The Board of Immigration Appeals provides little guidance on the application of the broad standard for asylum petitions, namely "reasonable fear of persecution." This lack of time for deliberation coupled with a very open-ended decision standard may amplify emotional influences. For replication purposes, we consider criminal sentencing by federal district judges in 1971–2012 (primarily 1998–2011), the only other large data base of comparable judicial decisions that we are aware of. Like the asylum data, sentencing cases are numerous and relatively homogeneous, and the outcomes are easy to classify. With hundreds or even thousands of similar cases per judge, we can thus construct fairly precise baseline approval or sentencing rates for each judge from the judge's own decision record. While we may like to believe that sentencing by federal district judges is not susceptible to influence by extraneous factors (due to higher quality of the federal judges, more time for deliberation, or the constraining effects of federal sentencing guidelines), in fact, district judges are susceptible to the same influences as asylum judges.

Our findings may have policy relevance since courts could impose a requirement that respondents are due free access to counsel in asylum cases. The positive effects of lawyers on outcomes

reports a far lower average grant rate of 16%.

[3] Eli Saslow, "In a crowded immigration court, seven minutes to decide a family's future," The Washington Post, 2/2/2014.

for respondents in observational studies are large, and we present evidence for one kind of mechanism — that the presence of a lawyer helps sharpen the analysis, removing arbitrary things from the outcomes, undercutting the negative effects of mood swings.

2. Data

The first empirical setting is U.S. asylum court decisions and the second is U.S. federal district court decisions.

2.1. Asylum Judges: Data Description and Institutional Context

The United States offers asylum to foreign nationals who can prove that (1) they have a well-founded fear of persecution in their own countries, and (2) their race, religion, nationality, political opinions, or membership in a particular social group is one central reason for the threatened persecution. Decisions to grant or deny asylum are potentially very high stakes for the asylum applicants. An applicant for asylum may reasonably fear imprisonment, torture, or death if forced to return to her home country (see Ramji-Nogales et al. [2007] for a more detailed description of the asylum adjudication process in the U.S.).

This article uses administrative data from 1993 to 2013 on U.S. refugee asylum cases considered in immigration courts. Judges hear two types of cases: affirmative cases (where the applicant seeks asylum on her own initiative) and defensive cases (where the applicant applies for asylum after being apprehended by the Department of Homeland Security (DHS)). Defensive cases are referred directly to the immigration courts while affirmative cases pass a first round of review by asylum officers in the lower level Asylum Offices. See Appendix A for more details regarding the asylum application process and defensive vs. affirmative applications.

The court proceeding at the immigration court level is adversarial and typically lasts several hours. Asylum seekers may be represented by an attorney at their own expense. A DHS attorney cross-examines the asylum applicant and argues before the judge that asylum is not warranted. Those that are denied asylum are ordered deported, although in some cases applicants may further appeal to the Board of Immigration Appeals.

Judges have a high degree of discretion in deciding case outcomes. They are subject to the supervision of the Attorney General, but otherwise exercise independent judgment and discretion in considering and determining the cases before them. This discretion is evidenced by the wide disparities in grant rates among judges associated with the same immigration court. Judges are appointed by the Attorney General and typically serve until retirement. Their base salaries are set by a federal pay scale and locality pay is capped at Level III of the Executive Schedule. In 2014, that rate was $167,000. Based upon conversations with the President of the National Association of Immigration Judges, no bonuses are granted.

We obtained the data directly from EOIR via a FOIA request (we also obtained a nearly identical data set via Transactional Records Access Clearinghouse (TRAC) and double-checked our results on those data). The data contains information on hearing dates, the completion date, whether the applicant was legally represented, whether the application was filed affirmatively or defensively (i.e., in defense of a removal proceeding), and the applicant's origin. We exclude non-asylum related immigration decisions and focus on applications for asylum, withholding of removal, or protection under the convention against torture (CAT). Applicants typically apply for all three types of asylum protection at the same time. As in Ramji-Nogales et al. [2007], when an individual has multiple decisions on the same day on these three applications, we use the decision on the asylum application because a grant of asylum allows the applicant all the benefits of a grant of withholding of removal or protection under the withholding-convention against torture while the reverse does not hold. The two categories are almost always ancillary to the asylum application, in which case they are not independent data points. There are only 22,000 independent withholding of removal and protection under the convention against torture applications, far fewer than the 434,000 asylum applications. We keep withholding of removal and protection under the convention against torture applications while only marginally increasing sample size, but only keep those that constituted independent applications.[4]

The main merit hearing is the hearing at which the case's substance is tried. Several practitioners have said that the judge will almost inevitably announce the final decision at the conclusion of the hearing. This is thus the relevant date for our purposes. However, the data does not explicitly flag the main merit hearing date. If the judge renders an oral decision at the hearing's conclusion, the main merit hearing date will coincide with the case completion date, which is in the data. We use the completion date, and drop from the data all cases for which the completion date does not coincide with a hearing date.[5] At this point, our data slims to 424,065 observations from the initial 456,686. For analyzing the impact of NFL games, 26,910 of the observations occur after a game and 88,456 are on Monday. The intersection of these data restrictions yields 22,294 observations. Over 89% of the decisions after an NFL game fall on Monday as opposed to the other days, so we restrict our baseline analysis to Mondays for the NFL analysis.[6] We later use all the Mondays to see if the wins increase grant rates or the losses decrease grant rates (or both) relative to Mondays that do not fall after a game.

Asylum seekers need to navigate complex legal challenges and those without access to

[4] We keep applications with a unique idncase idnproceeding.

[5] Sometimes, however, the judge reserves a written decision. In that case, the official completion date and the main hearing date do not coincide. Consistent with this, we find that this latter group of cases is more likely to involve a lawyer (94% vs. 90%), more likely to be a defensive case (46% vs. 38%), and—perhaps because the proportion of defensive cases is higher—less likely to result in a grant (36% vs. 39%). This introduces the theoretical possibility that the effects we observe are not true effects on the ultimate decision, but rather case composition effects as judges are more or less prone to reserve a written decision after a game was won. We have two replies to this. Firstly, the basic point would still go through: extraneous factors influence judicial decisions, even if the decision is procedural rather than substantive. Second, the number of decisions per day given our sample restriction is not systematically greater or smaller after wins. For the same reasons, and because they are reportedly very rare anyway, we are not worried that a greater or lesser rate of continuances after wins biases our results.

[6] The sample of Monday and Thursday night games is too small. Monday night games constitute a sample size one-tenth as large–and Thursday night games constitute a sample size of one-six hundredth as large–as the sample size of Sunday night games. Card and Dahl [2011] also exclude Monday and Thursday night games in their empirical analysis. The restricted sample has the advantage of observing judgments by a judge on the same day in different years, and judgments of different judges on the same day in a given year.

representation will have to represent themselves *pro se*. Many asylum seekers cannot afford to retain private counsel, which can be both costly and difficult to obtain, especially for detained asylum seekers, who cannot work to pay legal counsel fees. Free or low-cost legal representation is scarce in rural areas where detention centers are sometimes located, and federal funding restrictions limit the availability of legal services for asylum seekers (Ardalan [2014]).

2.2. Federal Sentencing Data

We obtain data on criminal sentencing by federal district judges from TRAC. Extensive description of these data is available elsewhere (Yang [2014]) and appendix L. In brief, federal district judges hear cases involving federal law and cases prosecuted by federal agencies. The roughly 700 judges are appointed for life by the U.S. President and confirmed by the Senate. The district court judgeships are among the most prestigious and revered judicial posts, only below that of the roughly 180 circuit court judgeships and the 9 on the Supreme Court. Thus, it becomes more hopeful that these judges would be more experienced and less susceptible to behavioral biases.

Criminal cases are prosecuted by the US Attorney, also politically appointed by the President. According to statistics from a recent study, 96% of defendants plead guilty, so there is no jury and only the sentence remains to be determined; 32% of cases have federal public defenders and another 21% have private counsel (McConnell and Rasul [2017]). The data span 1971 through 2012. For earlier years, we have only a selection of sentences, and very few before 1998. In total, there are approximately 900,000 cases.

The data contain information on prison sentences, probation sentences, fines, and the death penalty. The death penalty is exceedingly rare in federal cases (71 cases). Monetary fines are mostly very small relative to prison sentences. The median non-zero monetary fine is $2,000, and the 90th percentile is $15,000. We thus ignore them, and focus exclusively on prison sentences and probation.

The U.S. federal sentencing guidelines also help limit judicial discretion in sentencing. The guidelines specify a minimum and maximum sentence depending on offense severity and criminal history. However, judges can deviate from the guidelines if they find mitigating circumstances, such as family responsibilities, good work, prior rehabilitation, or diminished capacity.

Probation is another means with which a judge can mitigate discipline. Prior to the federal sentencing guidelines, probation would delay the imposition or execution of sentence. If a defendant violated a condition of probation, the court had the option to revoke probation and impose the prison sentence previously stayed. Probation as a means to suspend the sentence was abolished with the Sentencing Reform Act (1984), which recognized probation as a sentence in itself.

2.3. NFL Data

The article focuses on professional football because it is the most popular sport in the U.S.[7] We merged the asylum and sentencing data with NFL outcome data. As nearly all NFL games are played on Sundays, we dropped all other game days to keep the sample homogenous. We matched the courthouse of the judge to the NFL team most favored by the local community in 2013 according to Facebook likes.[8] In contrast, Card and Dahl [2011] assign all residents of a state to their "local" NFL team. They argue that "Weaker emotional cues presumably lead to attenuated estimates of the effect of wins versus losses. We suspect that our assignment procedure is likely to lead to a conservative assessment of the effect of emotional cues on family violence." They use 6 NFL teams (our sample includes 28). We do not know the personal preference of any given judge. We considered surveying the immigration judges, but figured that asking the judges about their sports preferences would generate a near zero response rate. While it is reasonable to guess that a judge who cares about football would follow the local team, he or she may not, and in fact may not care about football at all. The lack of separate information on judges' preferences also prevents disentangling whether sports influence decisions directly through the judges' mood, or through their environment or the other court house participants. However, we can use the sample of asylum cases that are resolved without lawyer representation. Moreover, the birth state of district judges (but not asylum judges) are available from the Federal Judiciary Center. More salient effects for judges born in the area of the courthouse would be suggestive that the effects are due to judge decision-making as opposed to the game or weather outcomes affecting other court participants such as lawyer behavior.

2.4. Weather

We use weather data from the National Weather Service. To combine the weather data with the courts data, we merge on date and location. We used rainfall, high winds, and snow.

This article does not claim that sports and weather are the main determinants of people's moods. But among the plausible influences on mood, they are ones we can actually measure for a large number of cases. The public has no access to data on judges' health status, family events, commuter traffic, etc. Other events, such as stock market crashes or terrorist attacks, are measurable and will likely have a much stronger effect on mood than weather or sports, but the sample size is (fortunately) much too small.

2.5. (Quasi-)Random Assignment

Obviously, the outcomes and characteristics of asylum cases do not influence NFL outcomes (neither directly nor through scheduling) or the weather. It is conceivable that case scheduling adjusts to NFL scheduling, outcomes, or the weather. For a number of reasons, however, this is

[7] This is confirmed by google search trends.
https://trends.google.com/trends/explore?date=all&geo=US&q=NFL,NBA,MLB,NHL

[8] Cf. http://www.facebook.com/notes/facebook-data-science/nfl-fans-on-facebook/10151298370823859. This method is reasonable since 94% of the data are between 1996-2013.

extremely unlikely. First, we have learned from conversations with practitioners that the main merit hearings in asylum cases are scheduled first-in first-out, leaving no role for discretionary adjustments. Second, even if there were such room, it seems implausible that NFL and the weather would enter the picture. In fact, many cases are scheduled so far in advance that not even the NFL schedule, let alone the result or the day's weather, would be known at the time of scheduling. The NFL schedule comes out in April[9], while asylum cases may be scheduled over a year in advance. Third, once scheduled, the main hearing date is essentially set in stone, and decisions are rendered on the spot in almost all cases. Finally, the article verifies empirically that cases heard after NFL wins are not statistically different on observable case characteristics (other than the grant decision) from cases heard after losses.

It is not easy to conceive of third factors that might influence both (unobserved) asylum case characteristics and NFL outcomes, let alone the weather. Perhaps cities that become wealthier attract (or cultivate) both a better football team and a more sophisticated set of asylum petitioners. The latter would be attracted by higher wages (although one might also think that economic migrants are unlikely to obtain asylum). The former would be attracted by the higher purchasing power and the concomitantly higher advertisement revenue. We account for this possibility by controlling flexibly for city time trends.

As prima facie evidence, we present regression discontinuity plots of the data. NFL outcomes make it easy to present a discontinuity graphically, especially for outcomes that are easy to classify like the granting of asylum. Figure 1 shows the grant rate plotted against the point differential in NFL games. We present a local polynomial regression overlaid on the raw data that is jittered.[10] Losses occur to the left of a 0 and wins occur to the right. An increase in the grant rate occurs when the court city's NFL team on the night before wins. The confidence intervals are wider to the edges since very few games have high realized point differentials. Interestingly, further away from 0, the effect is not so clear. This could be related to expectations.

[9] See, e.g., http://www.nfl.com/photoessays/0ap1000000161578.

[10] The grant rate is jittered to more clearly present the mass of data (grant rates are usually 0 or 1 for any given judge on a given day) and thus will occasionally appear outside [0,1].

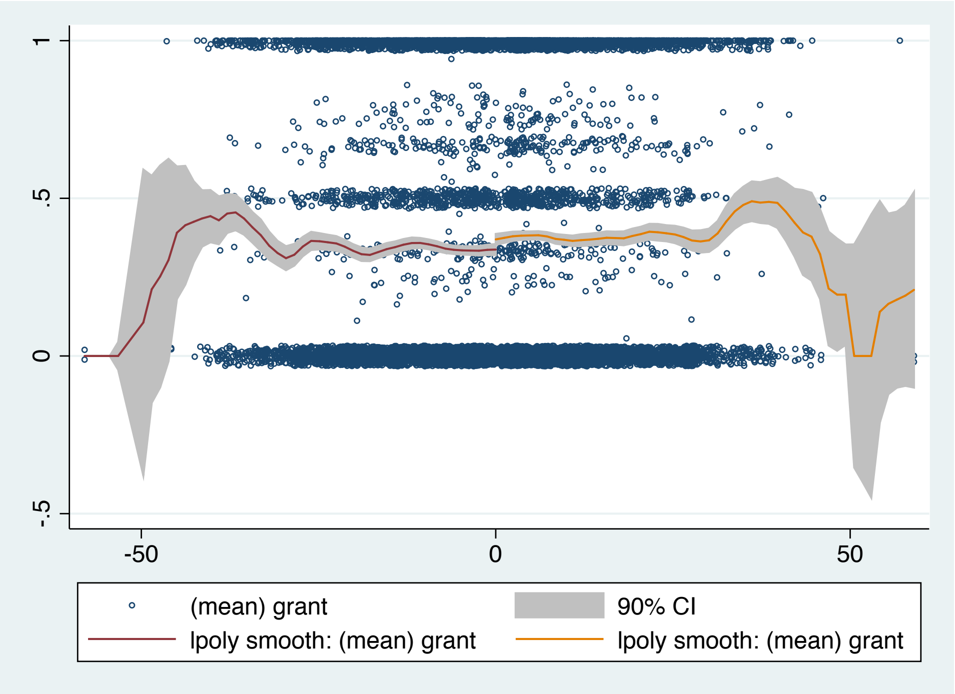


Figure 1: NFL & Asylum: Grant rates by point differences

Figure 2 shows the imprisonment rate and the probation sentence length plotted against the point differential in NFL games.

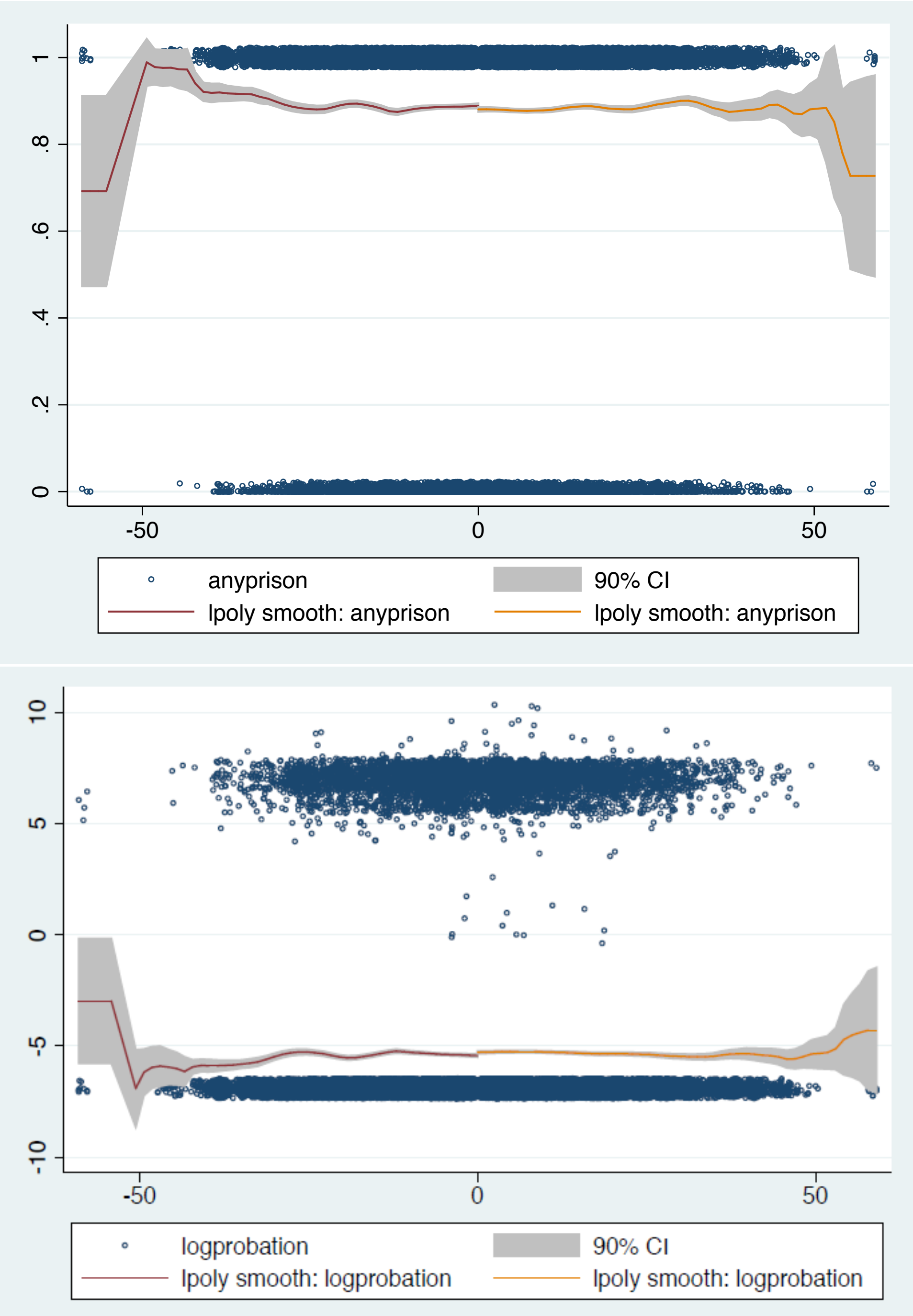


Figure 2: NFL & Sentencing: Prison and probation by point differences

3. Asylum Courts

3.1. Wins vs. Losses

We begin with the effect of NFL football wins vs. losses. For the reasons mentioned, we restrict the sample to asylum cases decided on Mondays after Sunday games. The sample is further collapsed to judge-city-day[11] and not every judge sits on a case on a Monday after NFL. We then present the effect of wins vs. no game and losses vs. no game, and finally, the effects of unexpected losses.

Table 1 estimates a fixed effects regression for applicants we, judges j, cities c, and decision date t of the following form:

$$Grantratio_{jct} = baserate_{jc} + \delta T_{ct} + \beta_1 X_{jct} + \beta_2 calendar_t + \epsilon_{jct}$$

$Grantratio$ is the ratio of grants to the number of decisions handed down by the judge in a given court house on a given day (that is, this reduces the dataset to at most one observation per judge per day). $Baserate_{jc}$ is a fixed effect for judge j sitting in city c. T_{ct} indicates the treatment and δ the coefficient of interest (e.g., win or loss). X_{jct} is a vector of average applicant covariates for the applicants who appeared before the judge in that court on that day. In particular, it contains whether the claim was defensive or affirmative, and whether the applicant was legally represented. We also include the fraction of applicants who were of the most frequent nationality.[12] $Calendar_t$ is a collection of calendar dummies for each week of the year (1-52) and for each NFL season between 1992 and 2013. ϵ_{jct} is an error term.

The case covariates X_{jct} are not required for identification. In fact, as already mentioned, we test that they are randomly distributed across treatment and control groups identified by T_{ct}. A separate issue is dependence of observations from the same city and, a fortiori, same judge. The standard way of dealing with dependence of observations is clustering. There are two levels at which cases are not independent, and they are not nested: the judge, and the city. we thus cluster either by city, judge, or both.[13] It hardly matters which way we cluster. In fact, we have found that the clustering surprisingly has only a small effect compared to no clustering.

[11] For example, if judge Smith granted four applications and denied one on 4/15/2013 in Newark and granted one in New York City, we would collapse this into two data points: one data point Smith-Newark-4/15/2013 with value 0.8, and one data point Smith-NYC-4/15/2013 with value 1.

[12] In the full sample, Chinese are over 20% of the applicants and by far the largest group. No subnational disaggregation is available. The next largest group is 7%.

[13] The results are similar clustering by judge, so we just present clustering by both city and judge.

Table 1: Main NFL Regressions

Dependent variable	Judge-City-Day Ratio of Granted Asylum						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Yesterday's NFL Win	0.019**	0.017*	0.018*	0.016**	0.014*	0.013*	0.013*
	(0.010)	(0.009)	(0.009)	(0.008)	(0.008)	(0.008)	(0.007)
Judge Fixed Effects	X	X					
City Fixed Effects		X					
JudgeXCity Fixed Effects			X	X	X	X	X
Season Fixed Effects				X			
JudgeXSeason Fixed Effects					X	X	X
Week Fixed Effects						X	X
Application controls							X
N	13504	13504	13504	13504	13504	13504	13504
R^2	0.21	0.23	0.23	0.26	0.44	0.44	0.46

Notes: Standard errors in parentheses (* p < 0.10; ** p < 0.05; *** p < 0.01) clustered at the judge and city level.

Controlling for application characteristics, the estimated effect of an NFL win is 1.3%. With only judge fixed effects, the estimated effect is slightly larger, namely 1 .9%. The difference between the two estimates is not statistically significant however. The estimated effect is stable with the gradual inclusion of controls, assuaging concerns of omitted variables.

Table 1 does not yet explicitly address the concern that applicant pools and NFL teams may develop in parallel. Table 2 explicitly addresses this possibility in two different ways. Models 1 and 2 include city-specific time trends, i.e., a separate time trend for each city. Here the coefficient stays at 1 .4%. However, the city-specific polynomial trend is rather crude.[14] A more flexible way to account for unobserved common trends is to match a decision to its nearest neighbor. That is, rather than imposing a particular polynomial model, we compare each decision to the closest decision by the same judge in the same city after the opposite game result. For example, if the city's team lost on weekend 47, we compare the decision on the following Monday to decisions after the nearest win(s): weekend 46 and 48, if any; if not, weekend 45 and 49, if any; and so on. Technically, this is a matching estimator (Abadie and Imbens [2006]). Model 3 requires an exact match on judge, city, and half-decade. Model 4 further requires the comparison be made to a match found within three months. This restriction hardly matters. The estimated effect is 1.9%. To address concerns of omitted variables another way, Appendix D presents the results of "placebo regressions" (balancing checks) using the application controls as the "outcome" variable.

[14] The city-specific seasonal trends include linear, quadratic, cubic, and quartic terms.

Table 2: NFL Regressions with flexible time controls

Estimation technique	OLS		Nearest-neighbor matching	
Dependent variable	Judge-City-Day Ratio of Granted Asylum			
	(1)	(2)	(3)	(4)
Yesterday's NFL Win	0.014*	0.014*	0.019*	0.019*
	(0.007)	(0.008)	(0.011)	(0.011)
Fixed Effects / Exact Match	JudgeXCity		JudgeXCityXHalfDecade	
Time control	City-specific trends		Match on date	
Time restriction				Within 3 months
Week Fixed Effects	X	X		
Season Fixed Effects	X	X		
Application controls	X	X		
N	13504	13504	7474	6832
Clustering	City	+Judge		
Number of clusters	56	56x340		

Notes: Standard errors in parentheses (* p < 0.10; ** p < 0.05; *** p < 0.01).

Appendix D also reports attenuation and anticipation estimates. The point estimates for the "effect" of Sunday night NFL games on the Friday before or on Tuesday decisions are very small with similar standard errors, which assuages concerns of the main estimated effects being due to statistical noise. Appendix E reports that no significant differences are found for whether the NFL team and the courthouse are in the same city[15], whether the game is played in the home city, or whether the game is a playoff game. These results suggest the mechanism is not due to fans attending the game.

3.2. Impacts of Wins vs. Impacts of Losses

Table 3 asks the separate question of the effects of (1) a loss, and (2) a win, compared to Mondays after non-game Sundays. We could then see if the effects of winning and losing are asymmetric. For example, if judges barely changed their decisions after a win as compared with an "untreated" Monday after non-game Sunday, but reacted negatively to losing (or vice versa), that might lead to a more complete understanding of the underlying psychology of mood. The estimation sample is the set of Mondays that occur through the NFL season. The results look largely due to losses. This is consistent with fans who experience loss aversion.

[15] There are 56 cities and 24 teams matched to asylum data.

Table 3: NFL Regressions with all Mondays

Dependent variable	Judge-City-Day Ratio of Granted Asylum	
	(1)	(2)
Yesterday's NFL Win	0.001	0.001
	(0.005)	(0.006)
Yesterday's NFL Loss	-0.014**	-0.014**
	(0.006)	(0.006)
JudgeXCity Fixed Effects	X	X
City-specific trends	X	X
Week Fixed Effects	X	X
Season Fixed Effects	X	X
Application controls	X	X
N	21468	21468
Clustering	City	+Judge
Number of clusters	56	56x340

Notes: Standard errors in parentheses (* p < 0.10; ** p < 0.05; *** p < 0.01).

3.3. Impacts of Upset Losses

Table 4 investigates the effects of unexpected losses. Card and Dahl [2011] regressed domestic violence on indicators for upset loss, close loss, upset win, predicted win, predicted close, and predicted loss. Eren and Mocan [2016] do the same with juvenile sentencing. Using the same specification, we find that in asylum decisions, an upset loss leads to 2.5% decline in the grant ratio. The point estimate of the effect of a loss when the game is predicted to be close is small with similar magnitude of standard errors. The estimated effects of an upset win are also small and not significantly different from 0. Card and Dahl [2011] and Eren and Mocan [2016] also report significant effects of upset losses and no significant impacts of close losses or upset wins. The coefficients associated with the range of the spread are significantly different from 0 and are potentially interesting, but less easily interpreted, since they may be correlated with other factors associated with the asylum grant rate. The coefficient is stable in more parsimonious models, which are presented in Appendix F.

Table 4: NFL Regressions with Mondays after NFL games

Dependent variable	Judge-City-Day Ratio of Granted Asylum	
	(1)	(2)
Loss X Predicted Win (Upset Loss)	-0.025**	-0.025**
	(0.011)	(0.012)
Loss X Predicted Close (Close Loss)	0.002	0.002
	(0.011)	(0.012)
Win X Predicted Loss (Upset Win)	0.002	0.002
	(0.011)	(0.013)
Predicted Win	0.053***	0.053***
	(0.012)	(0.012)
Predicted Close	0.027**	0.027**
	(0.012)	(0.013)
JudgeXCity Fixed Effects	X	X
City-specific trends	X	X
Week Fixed Effects	X	X
Season Fixed Effects	X	X
Application controls	X	X
N	21468	21468
Clustering	City	+Judge
Number of clusters	56	56x340

Notes: Standard errors in parentheses (* p < 0.10; ** p < 0.05; *** p < 0.01). Predicted Win indicates a point spread of -4 or less, Predicted Close indicates a point spread between -4 and 4 (exclusive), and Predicted Loss stands for a point spread of 4 or more. Predicted Loss is the omitted category.

3.4. Heterogeneity

This section examines whether the effects of NFL games are larger for unrepresented parties. This type of analysis would be suggestive that the effects are due to judge decision-making as opposed to the game outcomes affecting other court participants such as lawyer behavior.

The results are striking. NFL football games affect asylum cases more for unrepresented applicants. Table 5 shows that NFL outcomes affect the grant likelihood by 3.7% for defendants without lawyer representation. The effect of NFL win on unrepresented parties is statistically significant at the 1% level. When there is a lawyer, there is essentially no effect of the NFL outcome. The interaction term is statistically significant at the 10% level. Models 2 and 3 present the results for the sample with and without lawyers, which effectively fully interacts the controls with the presence of a lawyer. Appendix G shows the estimated coefficient is stable across model specifications, which assuages concerns of omitted variables that vary with the presence of a lawyer and the NFL win.

Table 5: Effect of NFL Outcomes by Lawyer Representation

Dependent variable	Granted Asylum		
	(1)	(2)	(3)
Yesterday's NFL Win	0.037***	0.006	0.027**
	(0.014)	(0.008)	(0.012)
Yesterday's NFL Win X Lawyer	-0.032*		
	(0.017)		
Lawyer	0.186***		
	(0.022)		
JudgeXCity Fixed Effects	X	X	X
City-specific trends	X	X	X
Week Fixed Effects	X	X	X
Season Fixed Effects	X	X	X
Application Controls	X	X	X
N	22282	20058	2224
Sample	All	With Lawyer	Without Lawyer

Note: Standard errors in parentheses (* p < 0.10; ** p < 0.05; *** p < 0.01). Standard errors are clustered by city. Observations are at the decision level.

This finding is consistent with the presence of lawyers overcoming the behavioral biases of judges, for example, by increasing the judge's attention to the case. It is also consistent with behavioral biases playing a larger role when judges are nearly indifferent for more disadvantaged applicants (Eren and Mocan [2016]). This leads us to suspect the NFL effects are not due to the lawyer behavior.

Table 6 reports a similar finding with unexpected outcomes. Upset losses affect the grant likelihood by 6.6% for defendants without representation. Interestingly, close losses also affect the grant likelihood for defendants without representation, by 4.6%. When there is a lawyer, there is essentially no effect of the NFL outcome. The interaction terms are statistically significant at the 5% level. Models 2 and 3 present the results for the sample with and without lawyers, to fully interact the controls with the presence of a lawyer.

Table 6: Effect of Unexpected NFL Outcomes by Lawyer Representation

Dependent variable	Granted Asylum		
	(1)	(2)	(3)
Loss X Predicted Win (Upset Loss)	-0.066***	-0.007	-0.067**
	(0.022)	(0.011)	(0.030)
Loss X Predicted Win (Upset Loss)	0.061**		
Lawyer	(0.023)		
Loss X Predicted Close (Close Loss)	-0.046**	0.008	-0.045**
	(0.022)	(0.011)	(0.021)
Loss X Predicted Close (Close Loss)	0.054**		
Lawyer	(0.024)		
Win X Predicted Loss (Upset Win)	-0.023	-0.001	-0.036
	(0.035)	(0.015)	(0.032)
Win X Predicted Loss (Upset Win)	0.020		
Lawyer	(0.036)		
JudgeXCity Fixed Effects	X	X	X
City-specific trends	X	X	X
Week Fixed Effects	X	X	X
Season Fixed Effects	X	X	X
Application Controls	X	X	X
N	22167	19948	2219
Sample	All	With Lawyer	Without Lawyer

Notes: Standard errors in parentheses (* p < 0.10; ** p < 0.05; *** p < 0.01). Standard errors are clustered by city. Observations are at the decision level. Predicted Win indicates a point spread of -4 or less, Predicted Close indicates a point spread between -4 and 4 (exclusive), and Predicted Loss stands for a point spread of 4 or more. Predicted Loss is the omitted category. All level terms, such as Predicted Win and Predicted Close, are included.

3.5. Weather

Table 7 looks at the effect of three types of bad weather on the day of the decision: rain, snow, and high winds. In each case, we have not only a dummy from the national weather service, but also a continuous variable measuring the intensity. We include city by week fixed effects so the weather variables are measured as a deviation from the norm for that week in that city. We include city by season fixed effects to control for trends in weather by city. We also include application controls, day of week fixed effects, and judge fixed effects. Thus, these effects capture intra-judge variation in asylum decisions. We again cluster the standard errors by city.

As can be seen, all three types of bad weather present reduce the grant rate. For example, the presence of snow reduces the grant rate by 1.0% and the effect is statistically significant at the 1% level. The presence of rain reduces grant rate by 0.2% but the effect is not statistically significant. The presence of high winds reduces grant rate by 2.3% and the effect is statistically significant at the 5% level. The intensity of bad weather does not have a statistically significant impact controlling for the presence of the bad weather. The F-test of joint significance rejects the null hypothesis of no effect in Columns 1, 3, and 4. Appendix H presents specifications to be comparable to the previous sections (judge by city and judge by season fixed effects). It hardly matters. Appendix H also presents a placebo regression with the lawyer representation. No effect is found for whether there is a lawyer.

Table 7: Judicial Decisions and Today's Weather

Dependent variable	Judge-City-Day Ratio of Granted Asylum			
	(1)	(2)	(3)	(4)
Snow present	-0.010***			-0.010***
	(0.003)			(0.004)
Snow amount in mm[1]	0.003			0.003
	(0.002)			(0.002)
Rain (may include freezing rain) present		-0.002		-0.001
		(0.002)		(0.002)
Precipitation in mm[1]		0.001		0.001
		(0.001)		(0.001)
Highwinds present			-0.023**	-0.024**
			(0.010)	(0.010)
Windspeed (tenths of meters per second)[1]			0.002	0.002
			(0.003)	(0.003)
F-Test of Joint Significance	0.020	0.372	0.074	0.005
Judge Fixed Effects	X	X	X	X
CityXWeek Fixed Effects	X	X	X	X
CityXSeason Fixed Effects	X	X	X	X
Application Controls	X	X	X	X
Day of Week Fixed Effects	X	X	X	X
N	239741	239741	239741	239741
R^2	0.29	0.29	0.29	0.29

Note: Standard errors in parentheses (* $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$). Standard errors are clustered by city. Observations are at the judge x day x city level. [1]Log of the underlying value+1.

We also checked if decisions for unrepresented parties are more affected by the weather. There are no statistically significant different weather effects for the two groups. One reason could be that the impact of weather is not overcome by a lawyers' presence. Another is that asylum applicants are affected by the weather—in a manner that does not happen with NFL games, which may be less relevant to asylum applicants—regardless of whether the lawyer is present.

Table 8: NFL and Sentencing Regressions with flexible time controls

Dependent variable	Any Prison		Probation Length[1]	
	(1)	(2)	(3)	(4)
Yesterday's NFL Win	-0.006**	-0.006**	0.050**	0.050**
	(0.003)	(0.003)	(0.020)	(0.020)
JudgeXCity Fixed Effects	X	X	X	X
DistrictXSeason Fixed Effects	X	X	X	X
Case controls	X	X	X	X
Week Fixed Effects	X	X	X	X
Yesterday's NFL Game	X	X	X	X
N	208,126	208,126	208,125	208,125
R^2	0.17	0.17	0.17	0.17
Clustering	District	+Judge	District	+Judge
Number of clusters	94	94x1344	94	94x1344

Notes: Standard errors in parentheses (* p < 0.10; ** p < 0.05; *** p < 0.01). Dependent variables are any prison sentence, log of probation sentence length, and whether the primary offense was for drugs (trafficking, communication, or possession). Case controls are whether or not the case was tried and—except in the drugs regression—the department of the offense classification. Regressions are restricted to Monday decisions and control for having an NFL game yesterday.

4. Sentencing Decisions

4.1. Linear Regression

A first question is if and to what extent the results generalize to other judicial settings. As already mentioned, immigration courts are rather special. They have an extremely high workload, the judges are not life-tenured judges, and the applicable legal standard is rather loose.

We thus ran similar tests with the federal sentencing decisions. The results are in Table 8. Two things are immediately apparent. First, the estimated coefficients are negative. That is, as with asylum decisions, judges appear to be, if anything, more lenient after a positive sports outcome. Controlling for defendant characteristics and flexible time trends, the estimated effect of an NFL win on imprisonment rates is a reduction of 0.6%, which is statistically significant at the 5% level. Judges also assign probation lengths that are 5% longer, also statistically significant at the 5% level. For federal felony convictions, probation is probably an indicator of lenience, as probation is often imposed as a substitute for imprisonment and there are functionally inexhaustible resources at the Federal level for incarceration.

We next run similar tests with expectations. The results are in Table 9. The estimated effect of an NFL upset loss on imprisonment rates is an increase of 1.6%, which is statistically significant at the 1% level. Judges also assign probation lengths that are 11% shorter, also statistically significant at the 1% level. The result is likely due to judges handing out more prison sentences and less probation sentences, since 99% of individuals with any prison sentence have zero probation sentence lengths, while 88% of individuals who do not receive a prison sentence have a positive probation sentence. In these regressions, case controls are whether or not the case was tried and the department of the offense classification. Regressions are restricted to Monday decisions after an NFL game.[16] Appendix I presents specifications to be comparable to the asylum analysis (judge by season fixed effects). It hardly matters. Appendix I also presents a placebo regression—no effect is found for whether the primary offense was for drugs. The point estimates are small and the standard errors similar in size to the first binary regression.

Next, we examine whether the effects of NFL games are larger for judges born in the area of the courthouse. This type of analysis would further support the inference that the effects are due to judge decision-making as opposed to the game outcomes affecting other court participants. The results are again striking. NFL football games affect judicial decisions for judges born in the state of the courthouse, but not those born in a different state. Table 10 shows The effects are statistically significant at the 1% level for judges born in the same state, but not statistically significant for judges born outside the state. Here we only present models that cluster standard errors at the district court level as the results are essentially identical also clustering at the judge level, as we see in the previous two tables.

[16] As before, the coefficients associated with the range of the spread are significantly different from 0 and are potentially interesting, but less easily interpreted, since they may be correlated with other factors associated with sentencing outcomes.

Table 9: NFL and Sentencing Regressions with flexible time controls

Dependent variable	Any Prison		Probation Length[1]	
	(1)	(2)	(3)	(4)
Loss X Predicted Win (Upset Loss)	0.016***	0.016***	-0.109***	-0.109***
	(0.005)	(0.005)	(0.039)	(0.039)
Loss X Predicted Close (Close Loss)	-0.002	-0.002	0.008	0.008
	(0.004)	(0.004)	(0.028)	(0.028)
Win X Predicted Loss (Upset Win)	-0.004	-0.004	0.050	0.050
	(0.008)	(0.009)	(0.047)	(0.047)
Predicted Win	-0.012***	-0.012***	0.071**	0.071**
	(0.005)	(0.005)	(0.033)	(0.033)
Predicted Close	-0.007	-0.007	0.059	0.059
	(0.005)	(0.005)	(0.037)	(0.037)
JudgeXCity Fixed Effects	X	X	X	X
DistrictXSeason Fixed Effects	X	X	X	X
Case controls	X	X	X	X
Week Fixed Effects	X	X	X	X
N	57037	57037	57036	57036
R^2	0.21	0.21	0.21	0.21
Clustering	District	+Judge	District	+Judge
Number of clusters	94	94x1344	94	94x1344

Notes: Standard errors in parentheses (* p < 0.10; ** p < 0.05; *** p < 0.01). Predicted Win indicates a point spread of -4 or less, Predicted Close indicates a point spread between -4 and 4 (exclusive), and Predicted Loss stands for a point spread of 4 or more. Predicted Loss is the omitted category. [1]Log of probation length in days+1.

Table 10: NFL and Sentencing Regressions by Judge Born-in-State

Dependent variable	Any Prison	Probation Length[1]	Any Prison	Probation Length[1]
	(1)	(2)	(3)	(4)
Loss X Predicted Win (Upset Loss)	0.020**	-0.145***	0.011	-0.042
	(0.008)	(0.051)	(0.008)	(0.060)
Loss X Predicted Close (Close Loss)	0.000	-0.004	-0.007	0.028
	(0.005)	(0.034)	(0.006)	(0.038)
Win X Predicted Loss (Upset Win)	-0.004	0.038	-0.003	0.074
	(0.010)	(0.063)	(0.011)	(0.065)
Predicted Win	-0.013	0.069	-0.010	0.058
	(0.008)	(0.053)	(0.008)	(0.059)
Predicted Close	-0.009	0.062	-0.002	0.045
	(0.007)	(0.047)	(0.008)	(0.051)
JudgeXCity Fixed Effects	X	X	X	X
DistrictXSeason Fixed Effects	X	X	X	X
Case controls	X	X	X	X
Week Fixed Effects	X	X	X	X
N	32654	32654	24383	24382
R^2	0.223	0.221	0.245	0.232
Clustering	District	District	District	District
Number of clusters	94	94	94	94
Sample	Born In State		Born Out-of-State	

Notes: Standard errors in parentheses (* $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$). Predicted Win indicates a point spread of -4 or less, Predicted Close indicates a point spread between -4 and 4 (exclusive), and Predicted Loss stands for a point spread of 4 or more. Predicted Loss is the omitted category. Columns 1-2 are limited to the judges born in the same state and Columns 3-4 are limited to judges born out of the state. [1]Log of probation length in days+1.

In the final replication of the asylum results, we assess the impact of bad weather on federal sentencing decisions. As can be seen in Table 11, the impact of bad weather on imprisonment is jointly significant at the 5% level. For example, the presence of rain increases imprisonment rate by 0.2% and the effect is statistically significant at the 10% level. The presence of high winds increases imprisonment rate by 0.9% and the effect is statistically significant at the 10% level. The F-test of joint significance rejects the null hypothesis of no effect on probation sentence length.

Table 11: Sentencing Decisions and Today's Weather

Dependent variable	Any Prison				Probation Length[2]
	(1)	(2)	(3)	(4)	(5)
Snow present	-0.004			-0.004	0.035
	(0.003)			(0.003)	(0.022)
Snow amount in mm[1]	-0.001			-0.001	0.002
	(0.001)			(0.001)	(0.007)
Rain (may include freezing rain) present		0.002**		0.002*	-0.011
		(0.001)		(0.001)	(0.007)
Precipitation in mm[1]		-0.0005*		-0.0004	0.003
		(0.0003)		(0.0003)	(0.002)
Highwinds present			0.008	0.009*	-0.043
			(0.005)	(0.005)	(0.037)
Windspeed (tenths of meters per second)[1]			0.001	0.001	-0.004
			(0.001)	(0.001)	(0.006)
F-Test of Joint Significance	0.114	0.108	0.199	0.022	0.093
Judge Fixed Effects	X	X	X	X	X
CityXWeek Fixed Effects	X	X	X	X	X
DistrictXSeason Fixed Effects	X	X	X	X	X
Case controls	X	X	X	X	X
Day of Week Fixed Effects	X	X	X	X	X
N	916129	916129	916129	916129	916129
R^2	0.16	0.16	0.16	0.16	0.15

Note: Standard errors in parentheses (* p < 0.10; ** p < 0.05; *** p < 0.01). Standard errors are clustered by city. [1]Log of the underlying value+1. [2]Log of probation length in days+1.

The effects of bad weather are weaker in the federal district courts than in the asylum courts, perhaps because the federal district court judges are more professionalized.

4.2. Random Forest Modeling

Random forests have become a highly competitive modeling tool, that performs well in comparison with many standard methods. They are popular, because (i) they can handle large numbers of variables with relatively small numbers of observations, (ii) can be applied to a wide range of prediction problems, even if they are nonlinear and involve complex high-order interaction effects, and (iii) produce variable importance measures for each predictor variable.[17]

In the section, we look past the recommended sentencing range and predict the sentence length within this range. We investigate sentence length percentile relative to the sentence guideline range as a dependent variable. This standardization allows us to look at where within a guideline range a sentence falls. The interpretation of this percentile measure is described in Table XII below.

	$< 0\%$	$0\% - 50\%$	$50\% - 100\%$	$> 100\%$
sentence length	below guideline minimum (rare)	between guideline minimum and midpoint	between guideline midpoint and maximum	above guideline maximum (rare)

Table 12: Interpretation of Range Percentile Measure

The details of the data sources and processing are given in Appendix K. To preview our results, Figure 3 shows marginal correlations between selected weather features and the sentence percentile suggesting a U-shape pattern between maximum temperature for the day and judicial decisions.

Such a dependency is also supported by Chen and Eagel [2017] who show that temperature is an important feature for asylum decisions. When it is too hot or too cold, asylum grant rates fall. Card and Dahl [2011] also report that domestic violence increases when the maximum temperature is over 80 degrees Farenheit.[18]

We compared the performance of three models, Random Forests (RF), Linear Regression and Gradient Boosting, and found that RF performed the best. We utilized parameter tuning to choose the best model from this hypothesis space. The optimal hyperparameters we found were min-samples-leaf = 9 and max-features = 0.6 (60% of features used in each node split).

Variable Importance in random forests We assume the reader is familiar with the basic construction of random forests which are averages of large numbers of individually grown regression/classification trees. The random nature stems from both "row and column

[17] Chen and Eagel [2017] report that extraneous factors, like weather, have roughly the same random forest importance weight as whether the asylum applicant has a lawyer or the applicant's nationality.

[18] As the temperature and mood link seem to be validated, in Appendix J we check the effect of NFL outcomes and snow, rain, and winds on twitter mood data measured daily for 1 year across 8 cities using data from Mislove et al. [2010]. NFL wins the day before improve mood, and the effect is statistically significant at the 10% level. The F-test of joint significance rejects the null hypothesis of no effect from bad weather. For example, the presence of high winds decreases mood, an impact that is statistically significantly at the 1% level. Baylis [2018] also documents a U-shape between temperature and sentiment measured in twitter.

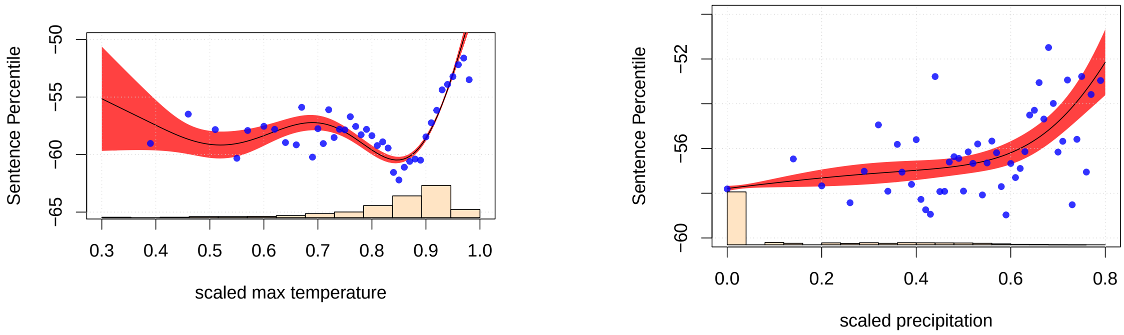


Figure 3: **left panel** We fit a generalized additive model to just the temperature to explore the marginal effects of the maximal temperature. Overlaid are the average values for the sentence percentile in bins of width 0.01, as long as the sample size is above $n = 500$. The histogram plays the equivalent of a rugplot and shows the distribution of the data on an arbitrary y scale. **right panel** Same for the maximum precipitation

subsampling": each tree is based on a random subset of the observations, and each split is based on a random subset of *mtry* candidate variables. The tuning parameter *mtry* – which for popular software implementations has the default $\lfloor p/3 \rfloor$ for regression and $\sqrt{p}$ for classification trees – can have profound effects on prediction quality as well as the to be introduced variable importance measures.

Our main focus in this paper is the CART algorithm Breiman et al. [1984], Breiman [2001] which chooses the split for each node such that maximum reduction in overall node impurity is achieved. Strobl et al. [2007a] et al. pointed out a bias of the CART algorithm towards categorical variables with different numbers of categories, or differing numbers of missing values. Recently, several authors [Loecher, 2020, Zhou and Hooker, 2021, Loecher, 2022] effectively eliminated the outlined bias inherent to the tree splitting procedure by including out-of-train samples in order to compute a debiased version of the MDI importance.

Alternatively, multiplicity-adjusted conditional tests could be used in the splitting process which avoid the known bias of the CART algorithm towards categorical variables with different numbers of categories, or differing numbers of missing values (Hothorn et al. [2006], Strobl et al. [2007a]). These so called *conditional inference* (CI) trees replace the CART bootstrap row sampling by sampling without-replacement of size $0.632 \cdot n$. In either case, 36.8% of the observations are (on average) not used for an individual tree; those *out of bag* (OOB) samples can serve as a validation set to estimate the test error, e.g.:

$$E\left(Y - \hat{Y}\right)^2 \approx OOB_{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{\hat{y}}_{i,OOB}\right)^2 \tag{1}$$

where $\bar{\hat{y}}_{i,OOB}$ is the average prediction for the ith observation from those trees for which this observation was OOB.

The splitting bias that was mentioned above also affects the originally proposed so-called Gini importance for classification and its analogue, average impurity reduction, for regression forests (Strobl et al. [2007b]). We will adopt the widely used alternative *reduction in MSE when permuting a variable* as a measure of variable importance defined as follows: VI $= OOB_{MSE,perm} - OOB_{MSE}$

An attempt at a theoretical foundation of variable importance for binary regression trees and forests is given in Ishwaran et al. [2007]. In related work (Ishwaran et al. [2008]), the authors point out that VI measures do not attempt to directly estimate the change in prediction error for a forest grown with and without the variable in question. We further note that the variable importance measure as defined above, has been shown to be closer to a measure of marginal importance rather than conveying the conditional effect of each variable (Archer and Kimes [2008], Strobl et al. [2008]). It can be shown that the permutation importance tests a joint hypothesis of independence between X_j and both Y and the remaining predictors $Z : X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p$: $H_0 : X_j \perp Y \wedge X_j \perp Z$. Hence a nonzero importance measure can be caused by a violation of either part: the independence of X_j and Y, or the independence of X_j and Z. The distinction between conditional and marginal influence is highly relevant for disentangling causal effects of (groups of) variables. For example, in our case we would like to make sure that the high variable imortances for weather and sports features are not simply due to geographic or temporal confounding. An alternative **conditional permutation scheme** is proposed in Strobl et al. [2008] which appears to mitigate the overestimation of the importance of correlated variables.

We refer the reader to Appendix L for the potential shortcomings of variable importance measures in random forests as well as a proposed solution to mitigate the confounding effect of correlated variables.

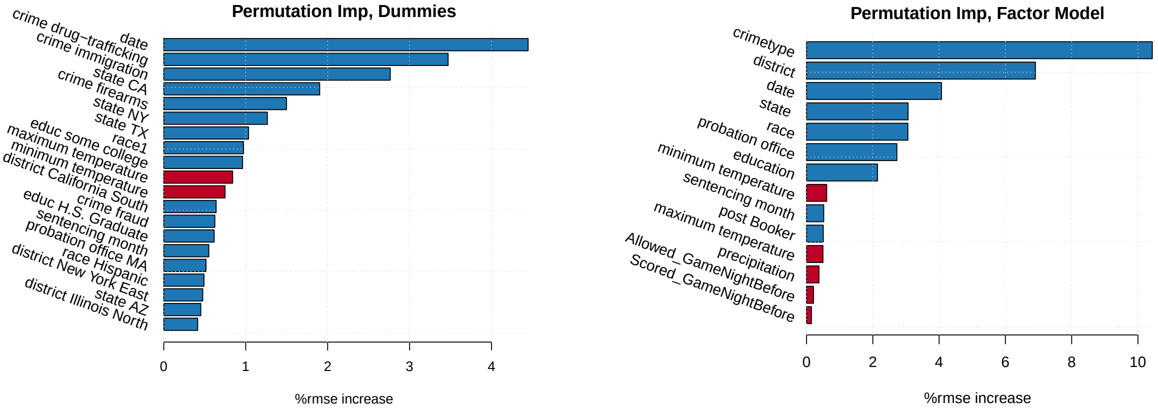


Figure 4: Normalized permutation importance – defined as percent increase in prediction-rmse when randomly shuffling a variable – for the original dataset without residualizing. We color code "unrelated" variables such as sports and weather features in red and the remaining variables in blue. **left panel** Model with dummy coding of factors. **right panel** Model with no dummifying of categorical variables.

Important Features Figure 4 displays the most important features based on a permutation scheme. While ”dummifying” categorical variables in linear models is well understood, its profound effects on model performance (Nick Dingwall [2016]) and key measures such as variable importance in tree based models are often overlooked. In fact, the majority of software implementations of random forests require dummy coding of categorical variables (Loecher [2018]) which makes benchmarking difficult. We feel that a fair and honest evaluation of the true impact of variables in machine learning needs to communicate both modeling approaches and their different interpretations. The left panel of Figure 4 evaluates the ”factor levels” individually while the right panel (no dummy coding) compares the overall contributions of the variables as a whole.

For the dummified model we found the most predictive feature the date of the sentencing decision included both as a continuous variable as well as a binary feature encoding the 2005 United States Supreme Court decision referred to as United States v. Booker (Wikipedia [2018a]), see also Figure 7 in Appendix K. For both models, location specific features capture high importance scores. For the left panel these would be specific states such as CA, NY, TX, AZ and districts 74, 7, 54 while the *factor model* yields high scores to state, district and location of probation office.

The most important feature related to the defendant was the crime type which is a reassuring sign that the judge is using case specific information in their decision. The ”dummy model” can be more specific w.r.t. the various types of crime and scores crimes involving drug traficking, immigration and firearms highest.

We find characteristics of the defendant that should not be important to be among the top 20 most predictive features, such as race and education level. It is important to further investigate whether these features truly influence the judge, as it would be unjust if they led to bias.

We found some weather features appear in our most predictive features. Temperature maximum and minimum were our 10th/11th as well as 8th/10th most predictive features, respectively.

We further found that some sports features, do in fact predict criminal sentence length to a small degree. It is worth noting that they are due to games that happened the prior day, not games that are going to happen. In fact we included the same sports features for the decision day but those accumulated no rmse reduction in predictions. That by itself supports a causal interpretation rather than it being a spurious correlation.

5. Conclusion

Our investigation into U.S. asylum and federal sentencing decisions over a three-decade span has highlighted a compelling dimension to the decision-making of bureaucrats: external factors, such as NFL game outcomes and weather conditions, can significantly impact judicial decisions. The data reveals that unrepresented parties are particularly susceptible to this phenomenon, and judges born in the home state of the respective NFL team are more likely to deliver harsher sentences after a team loss.

While we are unable to definitively ascertain whether these impacts are intentional or unconscious, their existence raises critical questions about the role of rationality and objectivity in decision-making. These effects demand a reevaluation of theoretical models that rely heavily on the assumption of rational, bias-free actors.

Our approach to the causal significance of these variables employs non-parametric partial correlation and residualizing methods. These techniques, grounded in the Frisch Waugh Lovell theorem and developments in double or orthogonal machine learning, unveil that seemingly unrelated variables like weather and sports events can indeed be detected as impacting judicial outcomes, raising the potential for automated methods to detect judicial indifference and personalize nudges of judges. Future research should explore these biases as potential indicators of systemic indifference and investigate interventions for improvement.

References

Alberto Abadie and Guido W. Imbens. Large Sample Properties of Matching Estimators for Average Treatment Effects. *Econometrica*, 74(1):235–267, January 2006.

Kellie J Archer and Ryan V Kimes. Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis*, 52(4):2249–2260, 2008.

Sabrineh Ardalan. Access to justice for asylum seekers: Developing an effective model of holistic asylum representation. *U. Mich. JL Reform*, 48:1001, 2014.

Anna Bassi, Riccardo Colacito, and Paolo Fulghieri. 'O Sole Mio: An Experimental Analysis of Weather and Risk Attitudes in Financial Decisions. *Review of Financial Studies*, 26(7): 1824–1852, February 2013.

Patrick Baylis. Temperature and temperament: Evidence from twitter. 2018.

L. Breiman. Random forests. *Machine Learning*, 45, 2001. doi: 10.1023/A:1010933404324. URL `https://doi.org/10.1023/A:1010933404324`.

Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984.

David V Budescu. Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression. *Psychological bulletin*, 114(3):542, 1993.

A. Colin Cameron, Jonah B. Gelbach, and Douglas L. Miller. Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, 90(3):414–427, August 2008.

Melanie Cao and Jason Wei. Stock market returns: A note on temperature anomaly. *Journal of Banking & Finance*, 29(6):1559–1573, 2005.

David Card and Gordon B. Dahl. Family violence and football: The effect of unexpected emotional cues on violent behavior. *The Quarterly Journal of Economics*, 126(1):103–143, 2011.

Daniel L. Chen and Jess Eagel. Can Machine Learning Help Predict the Outcome of Asylum Adjudications? *Artificial Intelligence and the Law*, March 2017. Accepted at ICAIL, TSE Working Paper No. 17-782.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/Debiased Machine Learning for Treatment and Causal Parameters. *ArXiv e-prints*, July 2016.

Marie Connolly. Some Like It Mild and Not Too Wet: The Influence of Weather on Subjective Well-Being. *Journal of Happiness Studies*, 14(2):457–473, April 2013.

Shai Danziger, Jonathan Levav, and Liora Avnaim-Pesso. Reply to Weinshall-Margel and Shapard: Extraneous factors in judicial decisions persist. *Proceedings of the National Academy of Sciences of the United States of America*, 108(42):E834, 2011a.

Shai Danziger, Jonathan Levav, and Liora Avnaim-Pesso. Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences of the United States of America*, 108(17):6889–6892, 2011b.

Richard B Darlington. Multiple regression in psychological research and practice. *Psychological bulletin*, 69(3):161, 1968.

Ronald Dworkin. *Law's Empire*. Harvard University Press, Cambridge, MA, 1986.

Alex Edmans, Diego Garcia, and Øyvind Norli. Sports sentiment and stock returns. *The Journal of Finance*, 62(4):1967–1998, 2007.

Ozkan Eren and Naci Mocan. Emotional judges and unlucky juveniles. Working paper, 2016.

George Everson. The Human Element in Judging. *Journal of the American Institute of Criminal Law and Criminology*, 10(1):90–99, May 1919.

Executive Office for Immigration Review. Office of the Chief Immigration Judge. "http://www.justice.gov/eoir/ocijinfo.htm", 2014.

Joshua B. Fischman and David S. Law. What Is Judicial Ideology, and How Should We Measure It? *Washington University Journal of Law and Policy*, 29:133–214, 2009.

Ulrike Grömping. Variable importance assessment in regression: linear regression versus random forest. *The American Statistician*, 63(4):308–319, 2009.

Ulrike Grömping. Variable importance in regression models. *Wiley Interdisciplinary Reviews: Computational Statistics*, 7(2):137–152, 2015.

Chris Guthrie, Jeffrey J. Rachlinski, and Andrew J. Wistrich. Inside the judicial mind. *Cornell Law Review*, 86(4):777–830, 2000.

Chris Guthrie, Jeffrey J. Rachlinski, and Andrew J. Wistrich. Blinking on the bench: How judges decide cases. *Cornell Law Review*, 93(1):1–44, 2007.

Andrew J. Healy, Neil Malhotra, and Cecilia Hyunjung Mo. Irrelevant events affect voters' evaluations of government performance. *Proceedings of the National Academy of Sciences of the United States of America*, 107(29):12804–12809, July 2010.

Anthony Heyes and Soodeh Saberian. Temperature and decisions: evidence from 207,000 court cases. *American Economic Journal: Applied Economics*, 2018.

Torsten Hothorn, Kurt Hornik, and Achim Zeileis. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, 15 (3):651–674, 2006.

Joseph C. Hutcheson, Jr. The Judgment Intuitive: The Function of the "Hunch" in Judicial Decision. *Cornell Law Review*, 14(3):274–288, April 1929.

Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer. Random survival forests. *The annals of applied statistics*, pages 841–860, 2008.

Hemant Ishwaran et al. Variable importance in binary regression trees and forests. *Electronic Journal of Statistics*, 1:519–537, 2007.

Jeff W Johnson and James M LeBreton. History and use of relative importance indices in organizational research. *Organizational Research Methods*, 7(3):238–257, 2004.

Duncan Kennedy. *A Critique of Adjudication*. Harvard University Press, Cambridge, MA, first paperback edition, 1998.

Markus Loecher. Categorical variables in trees i, 2018.

Markus Loecher. Unbiased variable importance for random forests. *Communications in Statistics - Theory and Methods*, 0(0):1–13, 2020. doi: 10.1080/03610926.2020.1764042.

Markus Loecher. Debiasing MDI feature importance and SHAP values in tree ensembles. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 114–129. Springer, 2022.

Brendon McConnell and Imran Rasul. Ethnicity, sentencing and 9/11. *Unpublished manuscript*, 2017.

Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J Niels Rosenquist. Pulse of the nation: Us mood throughout the day inferred from twitter. *Northeastern University*, 2010.

David B. Mustard. Racial, Ethnic, and Gender Disparities in Sentencing: Evidence from the U.S. Federal Courts. *The Journal of Law and Economics*, 44(1): 285–314, April 2001. ISSN 0022-2186. doi: 10.1086/320276. URL `https://www.journals.uchicago.edu/doi/abs/10.1086/320276`.

Chris Potts Nick Dingwall. Are categorical variables getting lost in your random forests?, 2016. URL `https://roamanalytics.com/2016/10/28/are-categorical-variables-getting-`

Appearing at a Master Calendar Hearing in Immigration Court. Political Asylum Immigration Representation Project, 98 North Washington Street, Ste. 106, Boston MA 02114, 2014.

Jeffrey J. Rachlinski, Sheri Lynn Johnson, Andrew J. Wistrich, and Chris Guthrie. Does Unconscious Racial Bias Affect Trial Judges? *Notre Dame Law Review*, 84:1195–1246, 2009.

Jeffrey J. Rachlinski, Andrew J. Wistrich, and Chris Guthrie. Altering Attention in Adjudication. *UCLA Law Review*, 60:1586–1618, 2013.

Jaya Ramji-Nogales, Andrew I. Schoenholtz, and Phillip G. Schrag. Refugee Roulette: Disparities in Asylum Adjudication. *Stanford Law Review*, 60(2):295–412, 2007.

Dan Simon. *In Doubt: The Psychology of the Criminal Justice Process*. Harvard University Press, Cambridge, MA, 2012.

Uri Simonsohn. Weather to go to college. *The Economic Journal*, 120(543):270–280, 2010. doi: 10.1111/j.1468-0297.2009.02296.x.

C. Strobl, A. L. Boulesteix, and T. Augustin. Unbiased split selection for classification trees based on the gini index. *Computational Statistics & Data Analysis*, 52, 2007a. doi: 10.1016/j.csda.2006.12.030. URL `https://doi.org/10.1016/j.csda.2006.12.030`.

C. Strobl, A. L. Boulesteix, A. Zeileis, and T. Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8, 2007b. doi: 10.1186/1471-2105-8-25. URL `https://doi.org/10.1186/1471-2105-8-25`.

Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC Bioinformatics*, 9(1):307, Jul 2008. ISSN 1471-2105. doi: 10.1186/1471-2105-9-307. URL `https://doi.org/10.1186/1471-2105-9-307`.

United States Courts. Court Role and Structure, a. URL `http://www.uscourts.gov/about-federal-courts/court-role-and-structure`.

United States Courts. Types of Cases, b. URL `http://www.uscourts.gov/about-federal-courts/types-cases`.

United States Sentencing Commission. United States Sentencing Commission. URL `https://www.ussc.gov/homepage`.

United States Sentencing Commission. United States Sentencing Commission. *Wikipedia*, June 2018. URL `https://en.wikipedia.org/w/index.php?title=United`$_{S}tates_{S}entencing_{C}ommission\&$
846979280.

Keren Weinshall-Margel and John Shapard. Overlooked factors in the analysis of parole decisions. *Proceedings of the National Academy of Sciences of the United States of America*, 108(42):E833, 2011.

Wikipedia. United States v Booker. *Wikipedia*, February 2018a. URL `https://en.wikipedia.org/w/index.php?title=United`$_{S}tates_{v}._{B}ooker\&oldid =$
827582160.

Wikipedia. Plea bargain. *Wikipedia*, June 2018b. URL https://en.wikipedia.org/w/index.php?title=Plea$_{b}argain\&oldid = 847752569$.

Wikipedia. College football. *Wikipedia*, June 2018c. URL https://en.wikipedia.org/w/index.php?title=College$_{f}ootball\&oldid = 845898433$.

Lee H Wurm and Sebastiano A Fisicaro. What residualizing predictors in regression analyses does (and what it does not do). *Journal of Memory and Language*, 72:37–48, 2014.

Cyril H Wyndham. Adaptation to heat and cold. *Physiology, environment, and man*, pages 177–204, 2013.

Crystal S. Yang. Have Interjudge Sentencing Disparities Increased in an Advisory Guidelines Regime? Evidence From Booker. *New York University Law Review*, 89(4):1268–1342, October 2014.

Zhengze Zhou and Giles Hooker. Unbiased measurement of feature importance in tree-based methods. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(2):1–21, 2021.

For Online Publication

A. Appendix A: Additional Background on Asylum Judges

Immigration Courts Overview

The immigration judges are part of the Executive Office for Immigration Review (EOIR), an agency of the Department of Justice Pol [2014]. At present, there are over 260 immigration judges in 59 immigration courts. In removal proceedings, immigration judges determine whether an individual from a foreign country (an alien) should be allowed to enter or remain in the United States or should be removed. Immigration judges are responsible for conducting formal court proceedings and act independently in deciding the matters before them. They also have jurisdiction to consider various forms of relief from removal. In a typical removal proceeding, the immigration judge may decide whether an alien is removable (formerly called deportable) or inadmissible under the law, then may consider whether that alien may avoid removal by accepting voluntary departure or by qualifying for asylum, cancellation of removal, adjustment of status, protection under the United Nations Convention Against Torture, or other forms of relief Executive Office for Immigration Review [2014].

Immigration Judges

The immigration judges are attorneys appointed by the Attorney General as administrative judges. They are subject to the supervision of the Attorney General, but otherwise exercise independent judgment and discretion in considering and determining the cases before them. See INA sec. 101(b)(4) (8 U.S.C. 1101(b)(4)); 8 CFR 1003.10(b), (d). Decisions of the immigration judges are subject to review by the Board pursuant to 8 CFR 1003.1(a)(1) and (d)(1); in turn, the Board's decisions can be reviewed by the Attorney General, as provided in 8 CFR 1003.1(g) and (h). Decisions of the Board and the Attorney General are subject to judicial review Executive Office for Immigration Review [2014]. Many previously worked as immigration lawyers or at the Immigration and Naturalization Service (INS) for some time before they were appointed.

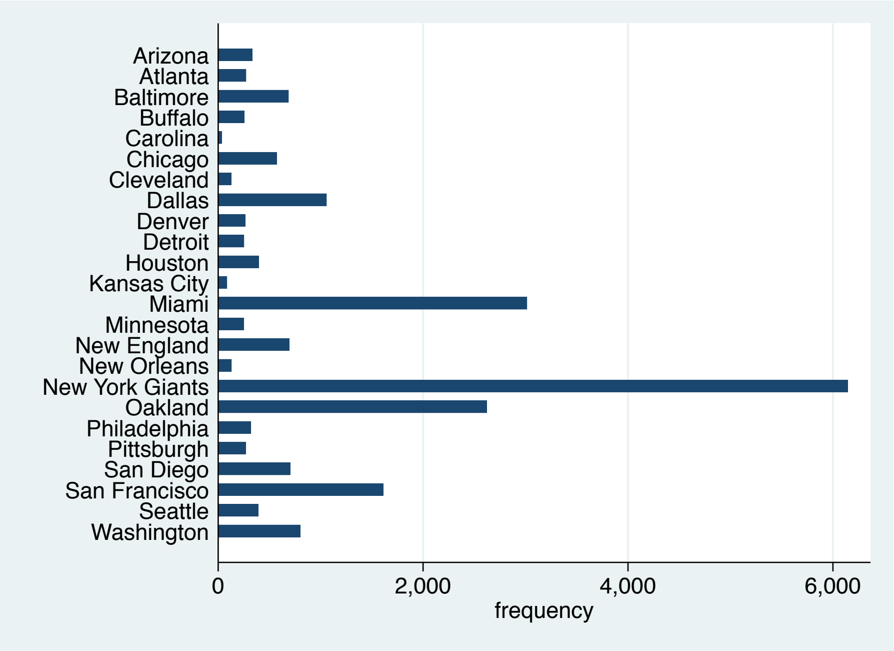
Proceedings before Immigration Courts

There are two ways an applicant arrives to the Immigration Court. First, the asylum seeker can affirmatively seek asylum by filing an application. In the event that the Asylum Office did not grant the asylum application[19] and referred it to Immigration Court, the asylum seeker can now pursue his or her asylum claim as a defense to removal in Immigration Court. Second, if the asylum seeker never filed for asylum with the Asylum Office but rather the government started removal proceedings against him or her for some other reason, he or she can now pursue an asylum case in Immigration Court Pol [2014]. This latter group is classified as defensive applicants and includes defendants picked up in immigration raids.

[19] For application at the Asylum Office, see chapters 14-26 of: http://immigrationequality.org/get-legal-help/our-legal-resources/immigration-equality-asylum-manual/preface-and-acknowledgements/

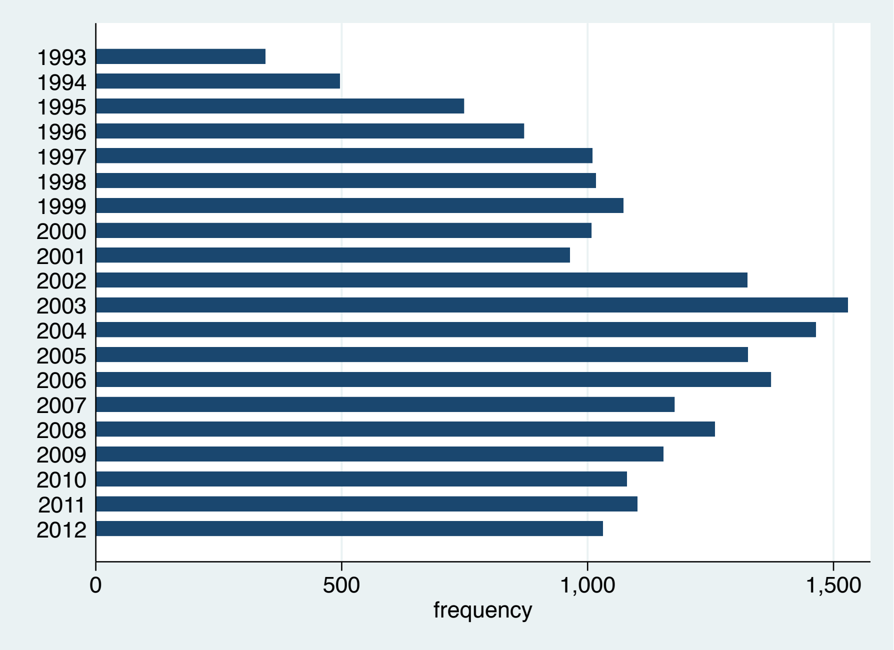
B. Appendix B: Distribution of Data

Appendix Figure 1: Distribution of Teams



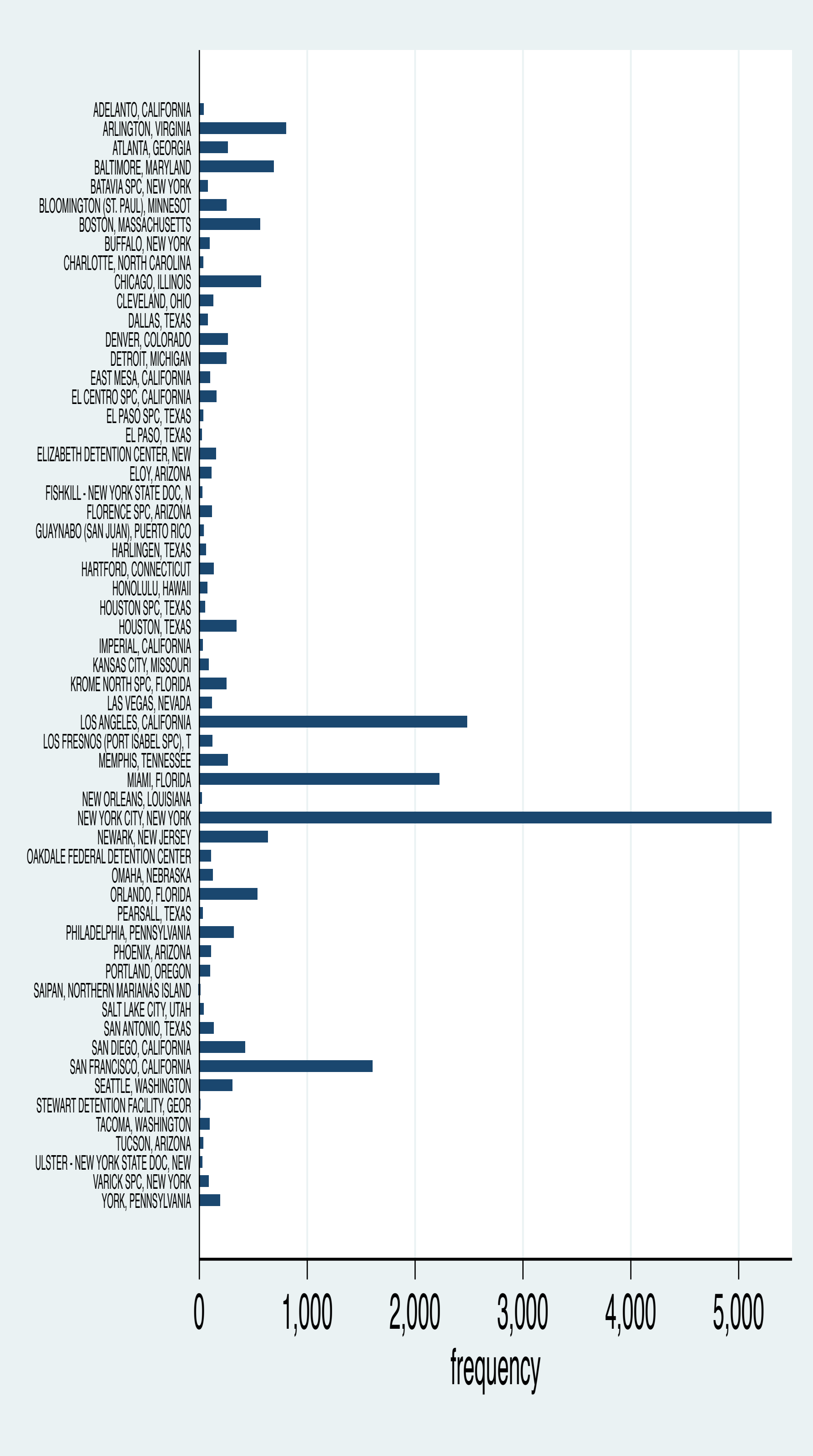
Notes: Asylum data restricted to Mondays after NFL games.

Appendix Figure 2: Distribution of Seasons



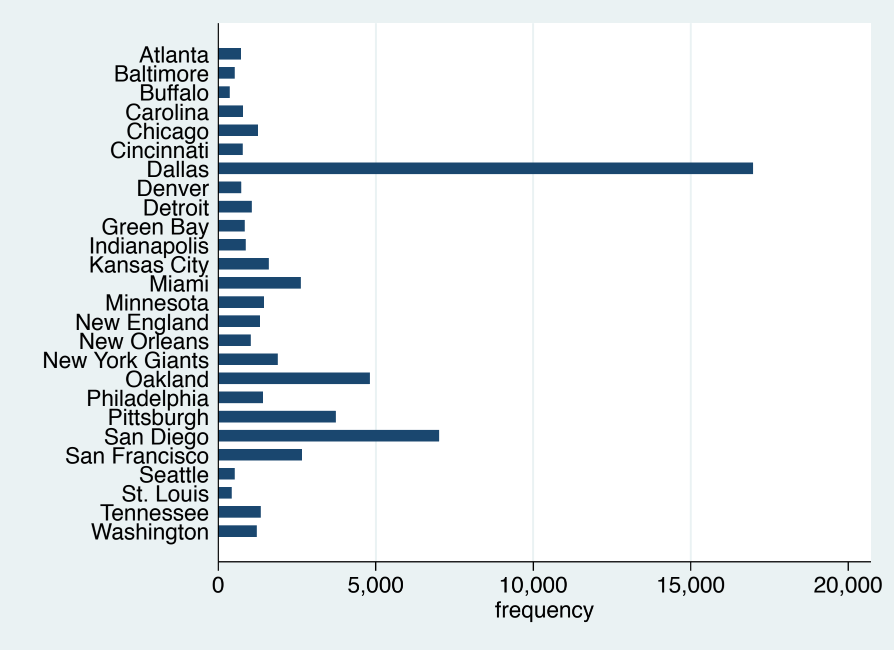
Notes: Asylum data restricted to Mondays after NFL games.

Appendix Figure 3: Distribution of Cities



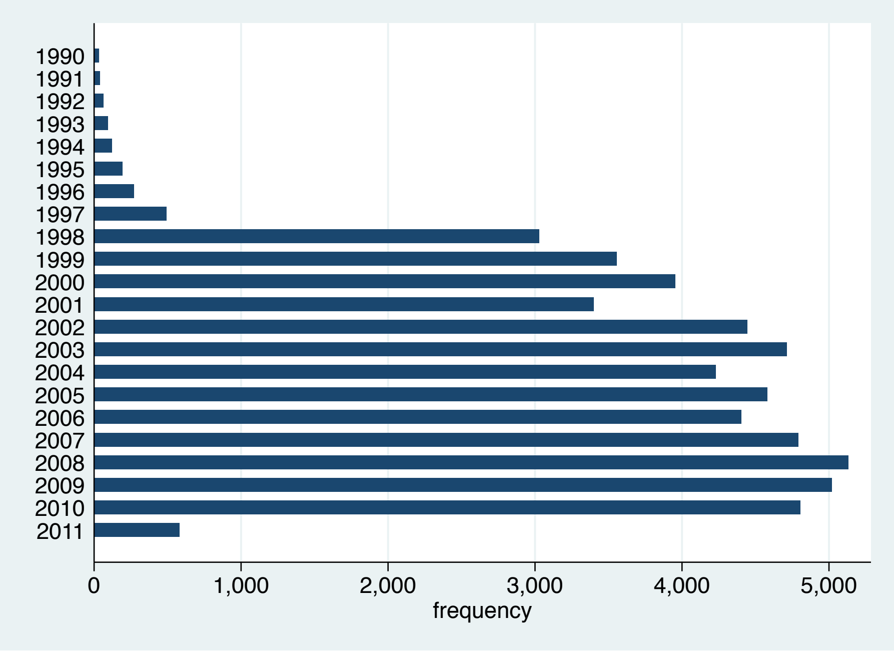
Notes: Asylum data restricted to Mondays after NFL games. Enlongated for readability.

Appendix Figure 4: Distribution of Teams



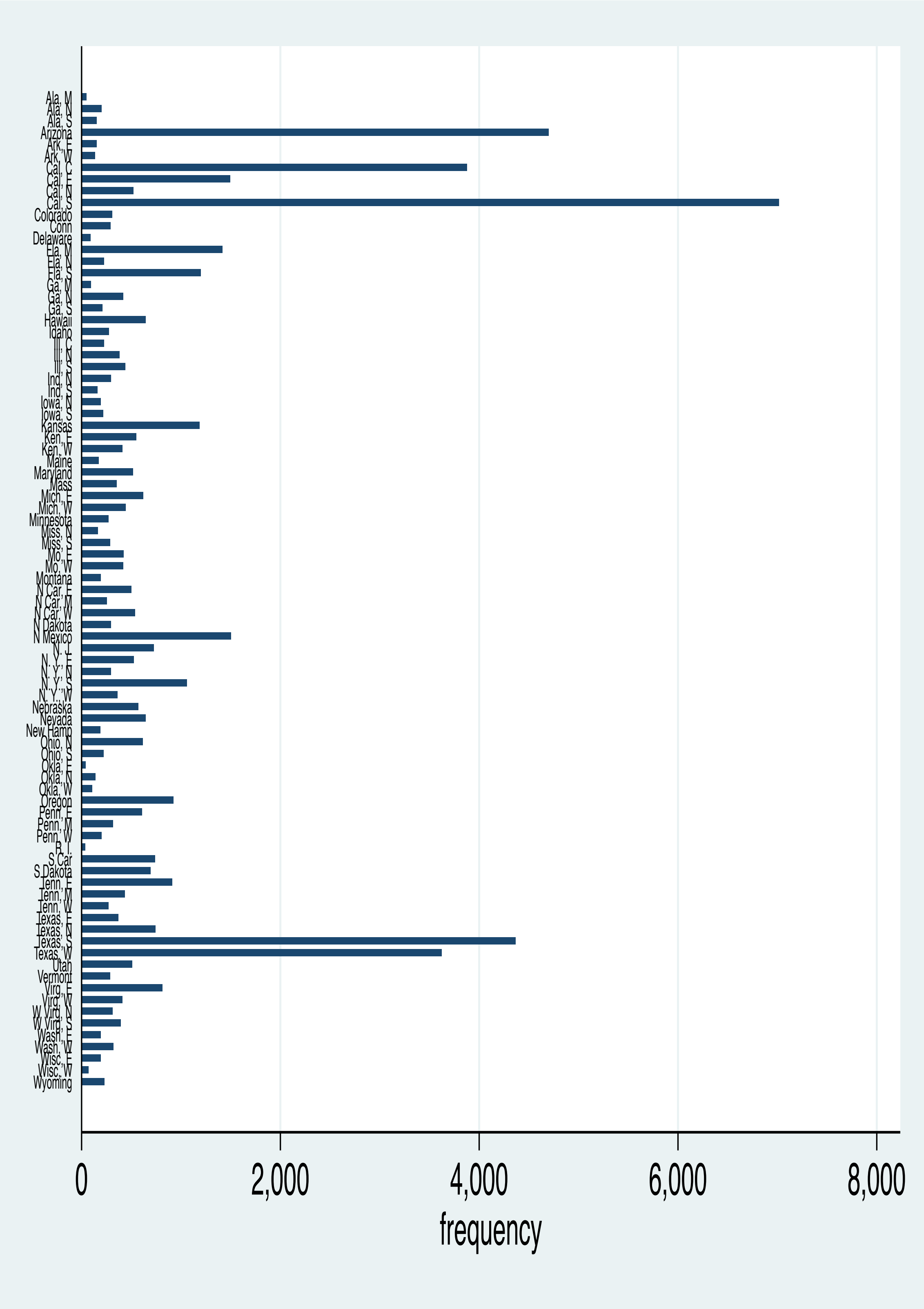
Notes: Sentencing data restricted to Mondays after NFL games.

Appendix Figure 5: Distribution of Seasons



Notes: Sentencing data restricted to Mondays after NFL games.

Appendix Figure 6: Distribution of Districts



Notes: Sentencing data restricted to Mondays after NFL games. Enlongated for readability.

C. Appendix C: Preliminary Bivariate Tests and Visualizations

This section presents some simple comparisons. Table A.1 compares mean grant rates on days after a win with grant rates after a loss. Looking at individual decisions (test 1), the average grant rate after a win is 3.7% higher than after a loss, or about 10% of the base grant rate. This difference is both economically large and, subject to the very important caveat in the next paragraph, highly statistically significant. Changing the unit of observation to individual judge-day grant rates (test 2) or city-day grant rates (test 3) barely changes this result. Similarly, grant rates are strongly positively correlated with wins, regardless of the level at which the data are pooled (Table A.2).

Appendix Table A.1: Differences in mean grant rates, by NFL win/loss

Level of aggregation	After	N	Mean	p-value (two-sided)
(1) Case	After loss	11101	0.371	
	After win	11193	0.408	
	Difference		-0.037	0.0000
(2) Judge-day	After loss	6676	0.345	
	After win	6795	0.379	
	Difference		-0.034	0.0000
(3) City-day	After loss	2596	0.291	
	After win	2620	0.318	
	Difference		-0.027	0.0099

Appendix Table A.2: Correlations between grant rates and NFL wins / win rates

Grant rates by		N	Correlation	p-value (two-sided)
Judge &	Day	13477	0.04	0.0000
	Season	3162	0.05	0.004
	Total	340	0.17	0.0013
City &	Day	5216	0.04	0.0099
	Season	845	0.10	0.0024
	Total	56	0.22	0.1105

To be sure, the simple statistical tests treat each case or ratio, as the case may be, as independent. In reality, however, observations from the same city and even more so from the same judge are subject to many of the same influences from unobserved factors. Moreover, the argument that NFL wins are randomly assigned to cases becomes tenuous over long time periods. As cities get richer, their football teams and asylum applicant pools may both become systematically stronger.

C.1. Power

Based on prior research on intra-judge differences, one important factor predicting case outcomes is the identity of the judge. There are 340 immigration judges in the asylum data set, compared to 1,268 district judges in the sentencing data set. Moreover, all asylum cases have the same binary potential outcome, while sentencing cases present vastly differing potential sentence ranges. To appreciate the demands on sample size, consider the following numbers. The asylum and sentencing data sets are the largest case data sets we are aware of, at present. The relevant subset of comparable decisions after a football game, however, only comprises 58,000 sentencing and 22,000 asylum decisions, respectively. A 1% treatment effect is thus 110 additional grant decisions in the treatment group relative to the control group. If we had only a tenth of the overall sample size, a mere 10 such additional decisions in the control and treatment group, respectively, could create the misleading appearance of a 1% treatment effect and would prevent any reasonable inference from such a smaller sample. We would not be able to claim with any certainty that the 1% estimated effect is a true effect or mere noise. Comparability of the underlying cases greatly facilitates bounding the probability of a chance result. Similarly, if we had at least a fairly good estimate of what the decisions should be absent the treatment, the actual difference would provide a fairly good estimate of the treatment effect.

Table A.3 presents summary statistics for all variables in the two datasets. Summary statistics for court cases and NFL outcomes are summarized over the data analysis sample restricted to Mondays after NFL games. The weather data is summarized for the entire data analysis frame. Appendix B presents distributions for cities, the teams, and over time.[20] Appendix C presents several motivating bivariate tests of the data. Those results offer intuition about clustering of standard errors described below.

[20] The percent of decisions occurring after games predicted to win is higher for sentencing. This is because district court sentencing decisions occur in regions and time periods more enthusiastic of teams predicted to win. For example, Dallas Cowboys is matched to 29% of the sentencing data but only 4% of the asylum data.

Appendix Table A.3: Summary Statistics

	Asylum		Sentencing	
	μ	σ^1	μ	σ^1
Grant	39%			
Defensive	39%			
Lawyer	90%			
Any Prison			88%	
Probation Length in Days^2			40	28
Drug			35%	
Trial			5%	
NFL Win	50%		51%	
Upset Loss	7%		8%	
Close Loss	25%		22%	
Upset Win	9%		7%	
Predicted Win	27%		32%	
Predicted Close	47%		43%	
Predicted Loss	26%		25%	
Snow present	4%		4%	
Snow amount in mm^2	40	50	40	51
Rain (may include freezing rain) present	38%		32%	
Precipitation in mm^2	92	139	91	147
Highwinds present	0.3%		0.4%	
Windspeed (tenths of meters per second)^2	40	18	35	16

Notes: ^{1}Standard deviations only presented for continuous variables. ^{2}Summarized for positive values.

D. Appendix D: Placebo Regressions - Balancing Checks and Attenuation/Anticipation Regressions

Table A.4 reports placebo regressions using the case covariates (i.e., lawyer, defensive, whether the defendant is from China). The point estimates are all small and the standard errors similar in size to the main regressions.

Appendix Table A.4: Placebo regressions using covariates as dependent variable

Estimation technique		OLS			Nearest-neighbor matching	
Dependent variable	Judge-City-Day Ratio of					
	Lawyer	Defensive	China	Lawyer	Defensive	China
	(1)	(2)	(3)	(4)	(5)	(6)
Yesterday's NFL Win	0.004	-0.008	0.005	0.005	0.0004	0.009
	(0.007)	(0.006)	(0.007)	(0.006)	(0.011)	(0.009)
Fixed Effects / Exact Match		JudgeXCity			JudgeXCityXHalfDecade	
Time control		City-specific trends			Match on date	
Week Fixed Effects	X	X	X			
Season Fixed Effects	X	X	X			
N	13508	13504	13508	7474	7473	7474
Clustering	City+Judge	City+Judge	City+Judge			
Number of clusters	56x340	56x340	56x340			

Appendix Table A.5: Attenuation and Anticipation regressions using Tuesday and Friday decisions

Estimation technique	OLS		Nearest-neighbor matching	
Dependent variable	Judge-City-Day Ratio of Granted Asylum on			
	Following Tuesday	Previous Friday	Following Tuesday	Previous Friday
	(1)	(2)	(3)	(4)
Sunday's NFL Win	0.009	0.001	-0.002	-0.016
	(0.008)	(0.009)	(0.010)	(0.014)
Fixed Effects / Exact Match	JudgeXCity		JudgeXCityXHalfDecade	
Time control	City-specific trends		Match on date	
Week Fixed Effects	X	X		
Season Fixed Effects	X	X		
Application controls	X	X		
N	14625	10248	8855	4749
Clustering	City+Judge	City+Judge	-	-
Number of clusters	56x340	56x340	-	-

Finally, we examine Tuesday decisions after and Friday decisions two days before Sunday NFL games. These regressions help assess the degree of attenuation or anticipation of the Sunday's NFL results. We report these results in Table A.5; the point estimates are all small and the standard errors similar in size as reported in Table 2.

E. Appendix E: Heterogeneity by Location or Time

Table A.6 checks for and finds no significant differences depending on the location of the game and time of year. Note that the coefficient for "Same city as NFL team" and "Playoffs" are less interpretable as they are associated with, inter alia, factors associated with grant rates that vary by region or time.

Appendix Table A.6: Effect of NFL Outcomes by Location or Time

Dependent variable	Judge-City-Day Ratio of Granted Asylum		
	(1)	(2)	(3)
Yesterday's NFL Win	0.015***	0.004	0.011
	(0.005)	(0.009)	(0.007)
Yesterday's NFL Win X	-0.012		
Same city as NFL team	(0.011)		
Same city as NFL team	0.023*		
	(0.012)		
Yesterday's NFL Win X		0.011	
NFL team plays at Home		(0.012)	
NFL team plays at Home		-0.004	
		(0.008)	
Yesterday's NFL Win X			-0.007
Playoffs			(0.032)
Playoffs			-0.014
			(0.023)
JudgeXCity Fixed Effects	X	X	X
City-specific trends	X	X	X
Week Fixed Effects	X	X	X
Season Fixed Effects	X	X	X
Application Controls	X	X	X
N	21346	21346	21346

Note: Standard errors in parentheses (* p < 0.10; ** p < 0.05; *** p < 0.01). Standard errors are clustered by city.

F. Appendix F: Unexpected NFL Outcomes

Table A.7 reports the effect of upset losses across specifications that vary the set of controls. The coefficient is stable across models.

Appendix Table A.7: NFL Regressions with Mondays after NFL games

Dependent variable	Judge-City-Day Ratio of Granted Asylum			
	(1)	(2)	(3)	(4)
Loss X Predicted Win (Upset Loss)	-0.025**	-0.029**	-0.032***	-0.031***
	(0.011)	(0.012)	(0.012)	(0.012)
Loss X Predicted Close (Close Loss)	-0.001	-0.008	-0.006	-0.003
	(0.011)	(0.014)	(0.012)	(0.012)
Win X Predicted Loss (Upset Win)	0.002	-0.011	-0.008	-0.005
	(0.013)	(0.012)	(0.012)	(0.012)
Predicted Win	0.054***	0.044**	0.042**	0.045**
	(0.012)	(0.018)	(0.018)	(0.018)
Predicted Close	0.032**	0.027**	0.024*	0.024**
	(0.012)	(0.012)	(0.012)	(0.012)
JudgeXCity Fixed Effects	X	X	X	X
Season Fixed Effects	X			
JudgeXSeason Fixed Effects		X	X	X
Week Fixed Effects			X	X
Application controls				X
N	13422	13422	13422	13418
R^2	0.27	0.47	0.47	0.48

Notes: Standard errors in parentheses (* p < 0.10; ** p < 0.05; *** p < 0.01). Predicted Win indicates a point spread of -4 or less, Predicted Close indicates a point spread between -4 and 4 (exclusive), and Predicted Loss stands for a point spread of 4 or more. Predicted Loss is the omitted category.

G. Appendix G: Lawyer Interactions

Table A.8 reports the effect of NFL outcomes by lawyer representation across specifications that vary the application controls that are also interacted with the presence of legal representation. The coefficients are stable across models.

Appendix Table A.8: Effect of NFL Outcomes by Lawyer Representation

Dependent variable	Granted Asylum				
	(1)	(2)	(3)	(4)	(5)
Yesterday's NFL Win	0.038***	0.035**	0.031**	0.035**	0.030**
	(0.014)	(0.014)	(0.015)	(0.014)	(0.014)
Yesterday's NFL Win X	-0.033*	-0.030*	-0.026	-0.030*	-0.024
Lawyer	(0.017)	(0.017)	(0.017)	(0.016)	(0.017)
JudgeXCity Fixed Effects	X	X	X	X	X
City-specific trends	X	X	X	X	X
Week Fixed Effects	X	X	X	X	X
Season Fixed Effects	X	X	X	X	X
Application Controls	X	X	X	X	X
Controls X Lawyer	Defensive	Origin	Week	Season	All
N	22282	22282	22282	22282	22282
Sample	All	All	All	All	All

Note: Standard errors in parentheses (* p < 0.10; ** p < 0.05; *** p < 0.01). Standard errors are clustered by city. Observations are at the decision level.

H. Appendix H: Weather Regressions

Table A.9 reports the effect of weather using a specification similar to the NFL analyses. The results are hardly affected.

Appendix Table A.9: Judicial Decisions and Today's Weather

Dependent variable	Judge-City-Day Ratio of Granted Asylum			
	(1)	(2)	(3)	(4)
Snow present	-0.010***			-0.010**
	(0.004)			(0.004)
Snow amount in mm[1]	0.002			0.002
	(0.002)			(0.002)
Rain (may include freezing rain)		-0.002		-0.002
present		(0.002)		(0.002)
Precipitation in mm[1]		0.001		0.001
		(0.001)		(0.001)
Highwinds present			-0.022***	-0.023**
			(0.008)	(0.009)
Windspeed (tenths of meters			0.001	0.002
per second)[1]			(0.003)	(0.003)
F-Test of Joint Significance	0.023	0.584	0.034	0.002
JudgeXCity Fixed Effects	X	X	X	X
JudgeXSeason Fixed Effects	X	X	X	X
Week Fixed Effects	X	X	X	X
Application controls	X	X	X	X
N	239253	239253	239253	239253
R^2	0.32	0.32	0.32	0.32

Note: Standard errors in parentheses (* p < 0.10; ** p < 0.05; *** p < 0.01). Standard errors are clustered by city. Observations are at the judge x day x city level. [1]Log of the underlying value+1.

Table A.10 presents a placebo regression. No effect is found for whether there is a lawyer.

Appendix Table A.10: Lawyer Representation and Today's Weather

Dependent variable	Judge-City-Day Ratio of Lawyer Representation			
	(1)	(2)	(3)	(4)
Snow present	-0.002			-0.002
	(0.003)			(0.003)
Snow amount in mm[1]	-0.000			-0.000
	(0.001)			(0.001)
Rain (may include freezing rain) present		-0.001		-0.002
		(0.001)		(0.001)
Precipitation in mm[1]		0.000		0.000
		(0.000)		(0.000)
Highwinds present			-0.005	-0.005
			(0.015)	(0.014)
Windspeed (tenths of meters per second)[1]			0.001	0.002
			(0.001)	(0.001)
F-Test of Joint Significance	0.648	0.579	0.390	0.693
Judge Fixed Effects	X	X	X	X
CityXWeek Fixed Effects	X	X	X	X
CityXSeason Fixed Effects	X	X	X	X
Application Controls	X	X	X	X
Day of Week Fixed Effects	X	X	X	X
N	239741	239741	239741	239741
R^2	0.220	0.220	0.220	0.220

Note: Standard errors in parentheses (* p < 0.10; ** p < 0.05; *** p < 0.01). Standard errors are clustered by city. Observations are at the judge x day x city level. [1]Log of the underlying value+1. Application controls omit lawyer representation.

I. Appendix I: Sentencing and NFL Regressions

Table A.11 reports the effect of NFL outcomes using a specification similar to the asylum analyses. The results are hardly affected.

Appendix Table A.11: NFL and Sentencing Regressions with alternative specifications

Dependent variable	Any Prison (1)	Probation Length[1] (3)
Loss X Predicted Win (Upset Loss)	0.014**	-0.096**
	(0.006)	(0.038)
Loss X Predicted Close (Close Loss)	-0.003	0.006
	(0.004)	(0.028)
Win X Predicted Loss (Upset Win)	-0.001	0.027
	(0.009)	(0.053)
Predicted Win	-0.011**	0.053
	(0.005)	(0.036)
Predicted Close	-0.007	0.059
	(0.006)	(0.039)
JudgeXCity Fixed Effects	X	X
JudgeXSeason Fixed Effects	X	X
Week Fixed Effects	X	X
Case controls	X	X
N	57037	57036
R^2	0.34	0.34

Notes: Standard errors in parentheses (* p < 0.10; ** p < 0.05; *** p < 0.01). Standard errors are clustered at the district level. Predicted Win indicates a point spread of -4 or less, Predicted Close indicates a point spread between -4 and 4 (exclusive), and Predicted Loss stands for a point spread of 4 or more. Predicted Loss is the omitted category. [1]Log of probation length in days+1.

J. Appendix J: Twitter Regressions

The final analysis considers the hypothesized mechanism using a proxy for mood. We examine the effect of NFL outcomes and bad weather on twitter mood data measured daily for 1 year across 8 cities using data from Mislove et al. [2010]. Table A.12 reports that NFL wins the day before improve mood, and the effect is statistically significant at the 10% level.[21] The F-test of joint significance rejects the null hypothesis of no effect from bad weather. For example, the presence of high winds decreases mood, an impact that is statistically significantly at the 1% level. Baylis [2018] also documents a U-shape between temperature and sentiment measured in twitter.

[21] Because of the small number of clusters, We also present robust standard errors without clustering. In addition, we execute wild bootstrap for the weaker result. Following Cameron et al. [2008] renders 95% confidence intervals between .006 and .09 for NFL wins.

Appendix Table A.12: Twitter Mood, NFL Outcomes, and Today's Weather

Dependent variable	Tweet Mood ($\mu = 6.4$)			
	(1)	(2)	(3)	(4)
Yesterday's NFL Win	0.047***	0.047*		
	(0.007)	(0.023)		
Snow present			0.015	0.015
			(0.016)	(0.019)
Snow amount in mm[1]			0.012**	0.012
			(0.005)	(0.007)
Rain (may include freezing rain) present			-0.041***	-0.041
			(0.004)	(0.023)
Precipitation in mm[1]			-0.006***	-0.006***
			(0.001)	(0.001)
Highwinds present			-0.098***	-0.098***
			(0.012)	(0.027)
Windspeed (tenths of meters per second)[1]			-0.030***	-0.030**
			(0.003)	(0.009)
F-Test of Joint Significance	0.00	0.08	0.00	0.00
City Fixed Effects	X	X	X	X
CityXWeek Fixed Effects			X	X
Day of Week Fixed Effects			X	X
N	1154	1154	25508	25508
R^2	0.21	0.21	0.29	0.29
Clustering	None	City	None	City
Number of clusters	-	8	-	8

Note: Standard errors in parentheses (* $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$). [1]Log of the underlying value+1.

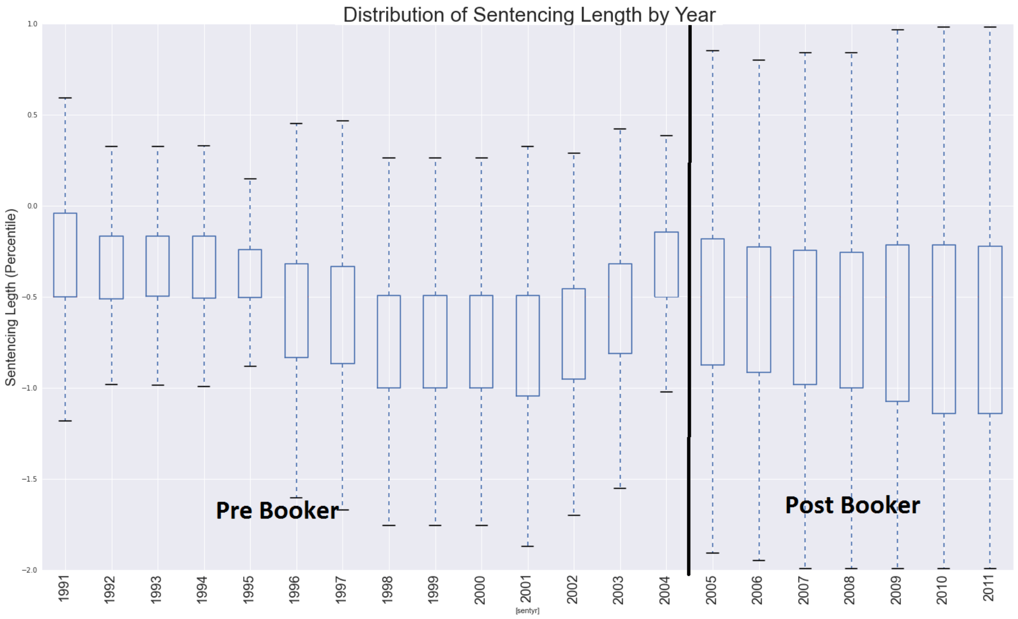
K. Appendix K: Sentencing Data

United States District Court

The United States District Courts (USDC) are the judicial backbone for hearing and sentencing federal crimes in the United States (United States Courts [a]). Federal crimes include illegal activity committed on federal land, crimes committed by or against federal employees in particular roles, matters involving federal government regulations (e.g., illegal immigration, federal tax fraud, counterfeiting), or crimes against the U.S. that occur outside of the United States, such as terrorism (United States Courts [b]). Among federal crimes, the most frequently heard cases involve immigration, drug trafficking, firearms, and fraud. Most frequently, the defendant in a case enters a plea agreement with the prosecutor, which is then approved of, or denied, by the judge (Wikipedia [2018b]). Otherwise, a sentencing trial is held and the judge determines the sentence for the criminal to serve: probation, federal prison, or both. In either situation, the judge has final say on the criminal sentence. There are 94 district courts in the United States. At least one district court is located in each state or U.S. territory. States that are large or have a large population have sub-state regional courts instead. The United States Sentencing Commission (USSC) (United States Sentencing Commission, 2018]), produces the sentencing guidelines for federal judges to use when they make their sentencing decisions. The judges are given a guideline range for the criminal sentence that is based upon the severity of the crime and the defendant's criminal history. Due to these guidelines, the largest factor determining sentence range is the criminal charges brought to the judge by the prosecutor.

The 2005 United States Supreme Court decision referred to as United States v. Booker (Wikipedia [2018a]) court decision determined that only prior convictions, facts admitted by the defendant, and facts proved to the jury beyond reasonable doubt could be used to extend the criminal sentence longer than the mandatory maximum. In other words, it introduced situations in which a judge could prescribe a sentence outside the sentencing range. We believe that this formal decision on opportunities to vary sentence length encouraged judges to change the way they made this determination. Interesting, while the U.S v. Booker case questioned the judge's right to increase the sentence length past the maximum guideline sentence, we saw an overall decrease in the length of sentence term relative to guideline range. Additionally, the range of minimum and maximum sentences becomes more extreme, as shown in Figure 7.

Discrepancies across choice of criminal charges do not fully explain these disparities. Judges are also known to, for example, give females a sentence nearer the guideline minimum, or prescribe criminal sentences outside of the guideline range for males (Mustard [2001]). This motivates our decision to focus on sentence length relative to the recommended guideline range. For the USDC, the Federal Sentencing Commission writes recommended sentence minimum and maximum terms to help ensure that convicts who committed similar crimes are charged with similar sentences. As can be seen in the lookup tables in United States Sentencing Commission, 2018], the severity of the crime and the criminal history of the convict are used to determine the appropriate sentence range. The judge then determines or approves a sentence length, frequently, but not necessarily within this range.



Appendix Figure 7: Trends in Sentencing Pre and Post US v. Booker.

Data Sources

The United States District Court Federal Sentencing data was made available by the Office of Research and Data in the United States Sentencing Commission. This data spanned federal court cases from 1992 – 2013. There are 35 features in this data, characterizing the defendant and crime. We keep 15 of these features due to their interpretability. For those models that cannot handle categorical fatures directly, dummy variables were created as needed for features including race/ethnicity, location and citizenship resulting in a total of 253 features. Our target variable was sentence length percentile relative to the range. We compute the value using standard normalization.

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

As our target variable was defined with the minimum and maximum sentence range, we dropped the minimum and maximum sentence range features when fitting our model to prevent data leakage.

Weather Data

In order to properly account for the weather in each district on a given day, we used a dataset originating from the NOAA (National Oceanic and Atmospheric Administration) database. This dataset consists of daily weather for 96 cities from 1992 – 2013. It includes over 90 features that depict various aspects of the weather conditions for each day. However, many of these features contain missing values, or are merely translations of other features. We chose to include only the following features: total daily sunshine and precipitation, maximum and minimum temperature, and average cloudiness.

Sports Data

Sports data available to us included data from MLB (Major League Baseball), NBA (National Basketball Association), NFL (National Football League), NHL (National Hockey League), and college football (CFB) for the years in which we had U.S. District Court Data. For the four professional leagues (MLB/NBA/NFL/NHL), there was an instance of each team in each game played (i.e. each game had two instances). While the features available were not identical across sports they were generally similar, and included information such as team name, field played on, score, and betting over/under. For the CFB data, there was one instance per game. Unlike the professional sports data, the CFB data is not as complete. This is understandable due to the organization of college football competitions. Teams typically play schools of the same size, budget, and quality of facilities (Wikipedia [2018c]). Due to this, some games played by smaller schools are not recorded. However, the games played by the Division I schools, the schools with the most developed football programs and likely the greatest regional following, are well represented. This data included team name, field played on, score, and so on. For each of the five sports datasets, we tabulated information about the each team per game on the same day as the trial, including the date, team name, whether a game occurred, and whether the game would be played at the home stadium, or away. We assumed that the judge would not know the result of the game before the end of the workday.

Our assumption is that the outcome of a game could influence a trial only by games played the day before and aligned the following features appropriately: whether the game occurred, whether the game would be played at the home stadium or away, the points scored by the team, the points scored by the opposing team, the score margin (difference between team's scores), and whether the team won or lost.

Data Processing

There were several challenges when pre-processing the sports data so that they could be organized into these dataframes. For example, for the score margin was not included in all data, and was calculated in these cases. The CFB data was organized differently from the professional sports data, so each instance of a game had to be split between the results per game per team. A lookup table between the team names and the district that would presumably be interested in that team was curated manually. For the lookup table to remain useful, several simplifying assumptions had to be made. In the first pass, each team was paired with the district where their home stadium was located. This meant that major cities such as Los Angeles, which is located in the Central California District Court district, were represented several times in the lookup table. New York City was challenging in that Brooklyn falls under the Eastern New York District Court, and the rest of New York City falls under the Southern New York District Court. In the majority of cases, New York City teams were represented by both districts, unless Brooklyn had its own team. After each team was paired with its ”hometown” district, we induced spatial spread in the professional sports data. First, in states that have several districts but only one team, the team was paired with all districts in that state. If a state had several districts and several teams, fandom maps based on Facebook likes were used to determine the more popular team in the ambiguous districts in that state. Finally, Boston teams were assigned to all New England districts, assuming homogeneity of fandom. If a district had no team and no obvious way to induce spread, it was not assigned any team (e.g., Guam, Puerto Rico, Montana, etc.). Due to the number of CFB teams, and assumptions about college football fan followings, we did not feel that spreading data outside of the district the school is located in was appropriate or desired. We choose not to use the betting over/under information included in the professional sports data, though that would be an interesting area of research worth pursuing. In the college football data, we choose not to include team ranking or whether the game was a special championship. An interesting future research aim would be to give a heavier weight to championship games and bowls, presuming that the lead up and results of the games would be more impactful on the community of fans invested in the game. Similarly, this information could be incorporated into the professional sports data.

Data Merge

To combine the weather data with district courts data, we merge on date and location. The features *city* and *courthouse* correspond to the location in the weather and district courts datasets, respectively. However, we found that the city names differ between the USDC and weather datasets. In other words, we found many courthouses for which there was no corresponding weather data. To avoid dropping criminal cases that do not have corresponding weather data, we created our own metadata to link courthouses in the district data to the nearest city in the weather data. Through this, we were able to precisely merge the two datasets without loss of information. The schema of this merge includes all district court features, along with weather features 0-4 in the weather table above.

To merge the sports data with the previously merged district court and weather data, we first dropped team name; we were interested to see if hometeam games affected the judges' sentence, rather than particular teams. For each of the sports dataframes described above, we merge over date and district. Each sport is represented separately. If no sports data was available for any day-district combination, the sports data fields were filled with zeros.

L. Appendix L: Variable Importance

We first introduce variable importance in the context of linear regression with p variables and n observations.

The concept of variable importance

Variable importance is not very well defined as a concept. Even for the case of a linear model with n observations, p variables and the standard $n >> p$ situation, there is no theoretically defined variable importance metric in the sense of a parametric quantity that a variable importance estimator should try to estimate (Grömping [2009]). In the absence of a clearly agreed true value, ad hoc proposals for empirical assessment of variable importance have been made, and desirability criteria for these have been formulated, for example, decomposition of R^2 into nonnegative contributions attributable to each regressor has been postulated (Grömping [2015]). An important distinction must be drawn between the two extremes of **marginal importance**, such as squared correlations versus **conditional measures**, e.g. squared standardized coefficients or sequential increase in R^2, as critically discussed, for example, by Darlington [1968].

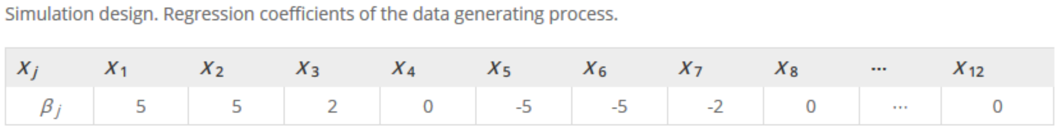
A recurring theme in the literature is that relative importance should balance out conditional and marginal considerations, a requirement brought forward by Budescu [1993] and later also by Johnson and LeBreton [2004].

Simulating data

For the sake of illustrating these concepts we generate a simple "linear" data set (no interactions, no nonlinearities)

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_{12} x_{12} \tag{2}$$

The predictor variables are sampled from a multivariate normal distribution $X_1, \ldots, X_{12} \sim N(0, \Sigma)$ where the covariance structure Σ is chosen such that all variables have unit variance $\sigma_{j,j} = 1$ and only the first four predictor variables are block-correlated with $\sigma_{j,j'} = 0.9$ for $j \neq j' \leq 4$, while the rest are independent with $\sigma_{j,j'} = 0$. Of the twelve predictor variables only six are influential, as indicated by their coefficients in Figure 8.

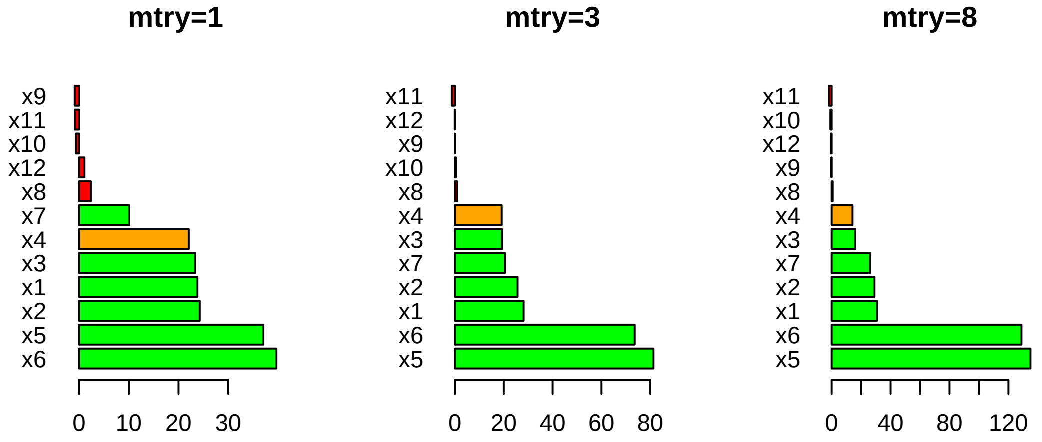


Simulation design. Regression coefficients of the data generating process.

x_j	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	...	x_{12}
β_j	5	5	2	0	-5	-5	-2	0	...	0

Appendix Figure 8: Population coefficients for simulated data in analogy to Strobl et al. [2008].

Notice the equal magnitudes (importance) of the set of coefficients $x_{1:4}$ and $x_{5:8}$ while only $x_{1:4}$ are correlated. The zero coefficients $x_{4,8:12}$ should get no weight which is confirmed by a linear regression. However, because of the imposed correlation structure, the variable x_4 might appear to be related to the dependent variable which could cause a high marginal variable importance. Generally speaking, for low values of $mtry$ we would expect the correlated predictors to serve as replacements of the truly influential ones. Figure 9 confirms this expectation and also the diminishing stand-in behavior of x_4 for increasing values of mtry.

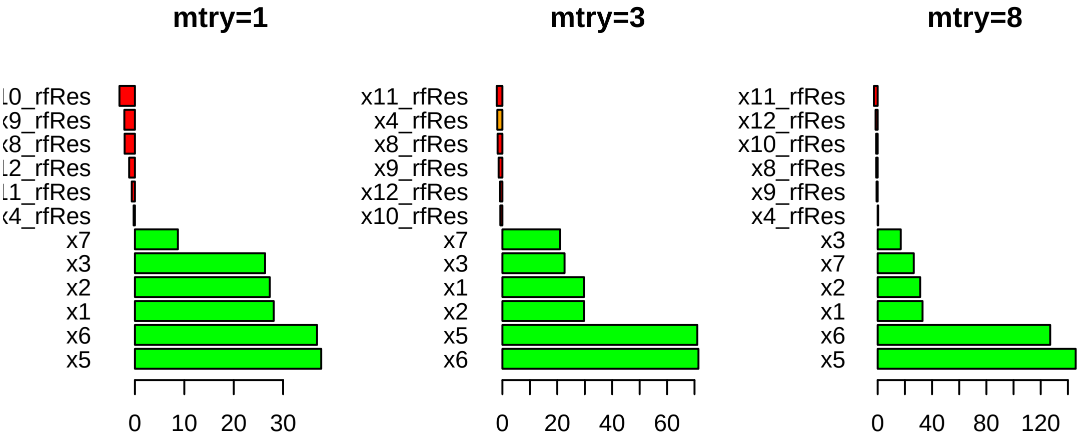


Appendix Figure 9: Permutation Variable Importance. The color coding is: green for truly nonzero coefficients, orange for correlated zero-value coefficients and red for all other $\beta_j = 0$. Note the slight negative values for the importance scores of the variabes with no predictive information.

We further observe that the correlation structure of $x_{1:4}$ dampens their individual VI scores: variables $x_{5:6}$ are consistently assigned a variable importance which is almost 4 times as high as the one for $x_{1:2}$.

Residualizing

The distinction between marginal and conditional variable importance in multiple linear regression is at the heart of the *ceteris paribus* interpretation of the estimated coefficients and covered in all introductory econometrics textbooks. The key insight we borrow is that the coefficient $\hat{\beta}_j$ does not change when we *residualize*, i.e. regress x_j on the remaining variables $x_{i \neq j}$. The effects of this type of residualizing in linear models are well understood though Wurm and Fisicaro [2014] highlights some undesirable effects. We extend the idea of residualizing in order to uncover the conditional effects of covariates to nonlinear models in analogy to the recently proposed concept of "Double Machine Learning" (Chernozhukov et al. [2016]). In particular, for each of the "seemingly unrelated" weather/sports variables $x_{i,SU}$ we train a random forest model using only the "appropriate features" as explanatory variables. We then replace the original $x_{i,SU}$ feature with the residuals $rfRes{-}x_{i,SU}$ from the respective auxiliary RF model. The main idea of this procedure is to remove existing correlations/dependencies between the two sets of variables and allow an interpretation of variable importance in the traditional sense of "controlling for XYZ". The results are promising for the simulated data. Figure 10 demonstrates that residualization appears to report conditional variable importances instead of marginal ones. We now apply the same idea to the court data in order to test whether the observed importance scores for seemingly unrelated variables in Figure 4 are robust under this "conditioning procedure".

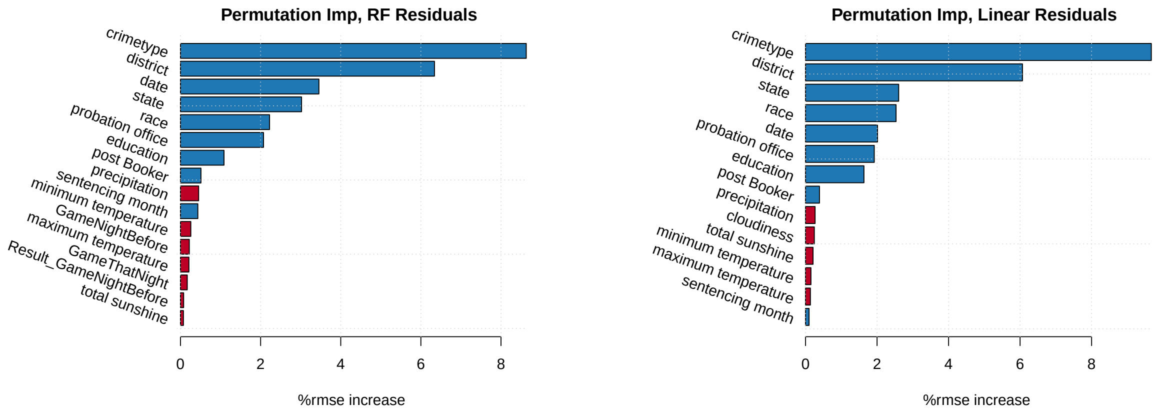


Appendix Figure 10: Permutation Variable Importance after replacing variables $x_{4,8:12}$ with their respective residuals from random forest models using $x_{1:3,5:7}$ as features. The color coding is as before.

Figure 11 shows the normalized permutation importance after each "unrelated" variable is replaced by the corresponding residuals from a random forest regression and confirms the robustness of the weather/sports feature influence.

Default Variable Importance

For completeness as well as a cautionary tale, in Figure12 we also provide the *mean decrease in impurity (or gini importance)* scores which happens to be the default choice in most software implementations of random forests. This mean decrease in impurity importance of a feature is computed as a (weighted) mean of the individual trees' improvement in the splitting criterion produced by each variable. A substantial shortcoming of this default measure is its evaluation on the in-bag samples which can lead to severe overfitting. It was also pointed out by Strobl et al. [2007b] that *the variable importance measures of Breiman's original Random Forest method ... are not reliable in situations where potential predictor variables vary in their scale of measurement or their number of categories.*



Appendix Figure 11: Normalized permutation importance after each "unrelated" variable is replaced by the corresponding residuals from a (*left panel*) random forest or (*right panel*) linear regression with the "appropriate" features as independent variables.The color coding is as in Figure 4

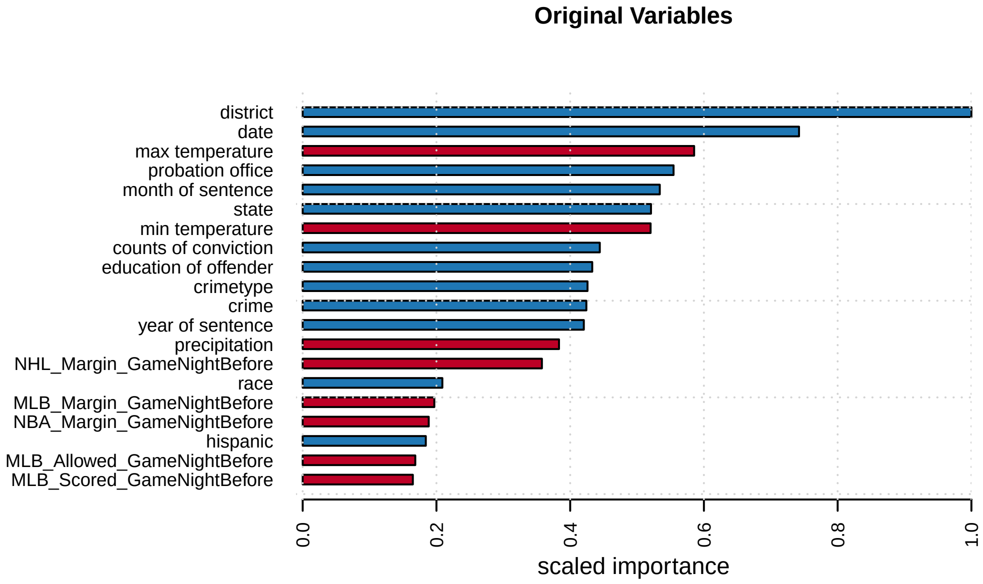
.

Null distribution of Importance Scores

It is not clear whether the observed smaller positive values of variable importance measures could be due to chance since there is no a well defined Null distribution for these scores. In order to identify cutoff values above which predictors would be considered to have a significant impact on model predictions, we have implemented a permutation test which generates a null distribution of importance scores for each predictor against which the observed importance scores are compared. This null distribution is created by randomly permuting the response variable (class assignments in a classification model, or independent continuous response in a regression model) among cases, running the same Random Forest model on the permuted data, and storing the resulting importance scores. (Note the computationally demanding nested permutations: an outer loop permuting the dependent variable and an inner loop shuffling the relevant predictors.) Under this procedure, a predictor that is not adding any significant information to the model will have an observed importance score that is similar to those generated by a random shuffling of the response, while a significant predictor will have an importance score much larger than the null. Significance p-values are then calculated as the fraction of replicates in the null distribution that are greater than or equal to the observed value.

Results: for the top 20 variables shown in Figure 4, the observed importance scores lie far above the extreme percentiles of the Null distributions, hence providing strong evidence for their significance.

We thank research assistants for contributions to the project and numerous colleagues with helpful comments. The paper benefits from data or code shared by Joshua Fischman, Jaya Ramji-Nogales, Andrew Schoenholtz, Phillip Schrag, and Crystal Yang, and from insights by Deborah Anker, Lenni Benson, Gordon Dahl, David Martin, Russell Wheeler, and especially Sue Long, and from Nora Barry, Laura Buchanan, Evelina Bakhturina, Erica Larson, Holger Spamann for early discussions and research assistance. Transactional Records Access Clearinghouse



Appendix Figure 12: Normalized *node purity importance* for the original dataset without residualizing. We color code "unrelated" variables such as sports and weather features in red and the remaining variables in blue.

(TRAC) at Syracuse University provided the sentencing data used in this paper as well as a separate copy of the asylum court data. This project was conducted while Chen received financial support from the European Research Council (No. 614708), Swiss National Science Foundation (No. 100018-152768), and Agence Nationale de la Recherche.

D Chen made contributions to Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, and Writing. M Loecher made significant contributions to Methodology, Validation, Software, Visualization, Writing of the original draft and revisions.