Cross Walking Bias: How Data Recombination Reifies Bias in Criminal Sentencing

Keywords: courts, court records, data cross walks, crime, bias

Extended Abstract

In the United States, the public has a constitutionally protected right to access court proceedings and to inspect court records. Established in a line of cases beginning with *Craig v. Harney* (1947) and notably including *Richmond Newspapers, Inc. v. Virginia* (1980), *Globe Newspaper v. Superior Court* (1982), and *Associated Press v. District Court* (1983), these rights are broadly reified in courts across the country. Despite this, court records largely remain opaque and inaccessible. Systems like the Public Access to Court Electronic Records (PACER) that ostensibly should solve this problem remain dated, difficult to use, and expensive. Even purchasing 1 year of data for a single state can cost up to $100,000 US dollars (Pah et al. 2019), casting doubt on whether these public records are meaningfully accessible to the general public.

That is where initiatives like the Systematic Content Analysis of Legitimation EventS (SCALES) come in. SCALES has undertaken the problem of improving the transparency of the federal courts by building an AI powered data platform that makes PACER court records – via clear visualizations and data downloads – accessible to everyone (SCALES 2023). However, much like any data innovation that enters the public space, SCALES and its underlying PACER data do not exist in a vacuum. Instead, they are part of a data universe where scholars and scientists are merging data together to generate new insights. Importantly, data producers have a responsibility to understand how their data exists within this larger universe and, we argue, should be prepared to offer findings on how their data can be responsibly cross walked with other data.

In this article we expressly consider how SCALES data insights can (or cannot) be combined with another important source of public criminal court data: The United States Sentencing Commission (USSC) data. USSC publicly provides information on sentences imposed in federal criminal cases involving individuals, but with a limited number of variables. Several scholarly attempts have been made to combine data from PACER with data from USSC to gain additional analytic leverage. Notably, Schanzenbach and Tiller (2007), Yang (2014), and Ciocanel et al. (2020) all attempted to merge USSC data to PACER data. Taking Ciocanel et al. (2020) as the most recent example, they began with 1,265,688 USSC records. They were able to match 804,128 of them with intermediary data. They were then able to match 524,393 of these records to PACER via district and docket number. That means the total data loss was 741,295 records or approximately 59% of data.

We conduct our own investigation of the best possible match between PACER and USSC data to see if this 59% data loss (and the data loss fond in other studies) systematically matters for conclusions about criminal justice sentencing. We conclude that there is systematic data loss particularly about Black and Hispanic defendants that can serve to obfuscate bias in federal sentencing by virtue of removing the most common offense types and most prevalent minority groups from the data altogether.

We find that the best possible match for PACER and USSC data includes the following variables: sentence month, sentence year, total months sentence, number of total

counts, probation, district, and the amount of restitution. In a test of a randomly chosen year (2016) we found that 34,258 cases could be identified uniquely using these variables, but that 32,615 could not. Even if those 32,615 cases had only one duplicate, our chances of guessing the correct answer are only 50%. As shown in Figure 1, we find that duplication varies by crime type and racial group. Drug sentences (trafficking and possession), firearm sentences, and immigration sentences had the most duplicates and these 4 categories make up 73.02% of crimes reported in the USSC overall. We also find that Black or African American and Hispanic individuals are disproportionately overrepresented in these categories.

That means we cannot simply drop duplicates or unmatchable data without contextualizing use cases. If we do simply drop the unmatchable data we are 1) systematically removing the most common offenses from the data, 2) systematically removing Black and Hispanic defendants from the data, 3) making it impossible to accurately measure racial disparity and other form of inequity. These three problems can persist even if the overall percentage of data loss seems acceptable on its face.

Rather than simply discouraging cross-walking these two important data sources, our work instead aims to elucidate the more limited list of use cases for which PACER and USSC data can be confidently merged. In keeping with this aim, we conclude our article by introducing several such use cases and their substantive findings. Also, to demonstrate the magnitude of risk of 'bad crosswalks,' we present a negative use case that quantifies the extent to which bad data combination can obscure inequality. The impact of this paper is not only a cross walked dataset for other scholars, an improvement on matching procedures, and illuminated use cases, but also rests on a larger argument about the role of scholars in ethical data production and combination.

References

Ciocanel, Maria-Veronica, Chad M. Topaz, Rebecca Santorella, Shilad Sen, Christian Michael Smith, and Adam Hufstetler. "Justfair: Judicial system transparency through federal archive inferred records." *Plos one* 15, no. 10 (2020): e0241381.

Pah, Adam R., David L. Schwartz, Sarath Sanga, Zachary D. Clopton, Peter DiCola, Rachel Davis Mersey, Charlotte S. Alexander, Kristian J. Hammond, and Luís A. Nunes Amaral. "How to build a more open justice system." *Science* 369, no. 6500 (2020): 134-136.

Schanzenbach, Max M., and Emerson H. Tiller. "Strategic judging under the US sentencing guidelines: Positive political theory and evidence." *The Journal of Law, Economics, & Organization* 23, no. 1 (2007): 24-56.

Systematic Content Analysis of Litigation EventS (2023). Available at: https://scales-okn.org/

Yang, Crystal S. "Have interjudge sentencing disparities increased in an advisory guidelines regime-evidence from Booker." *NYUL Rev.* 89 (2014): 1268.

Cases Cited

Associated Press v. U.S. District Court, 705 F.2d 1143, 1148 (9th Cir. 1983).
Craig v. Harney. 331 U.S. 367 (1947).
Globe Newspaper Co. v. Superior Court, 457 U.S. 596, 601 (1982).
Richmond Newspapers, Inc. v. Virginia, 448 U.S. 555, 556–57 (1980).

Figure 1. Duplications Due to Match Constraints

Two panels depicting different duplication levels across the data.

Panel A. Duplicated Crime Categories by Race

Offense by Race					
	White	Black	Hispanic	Other	Total
murder	11	17	18	26	72
manslaughter	2	1	1	58	62
kidnapping	13	12	29	8	62
sexual abuse	213	209	93	134	649
assault	153	110	125	388	776
robbery	269	279	78	22	648
arson	24	5	3	7	39
drugs (trafficking)	4,255	4,609	9,551	607	19,022
drugs (possession)	182	132	869	51	1,234
firearms	1,927	4,160	1,707	246	8,040
immigration	371	285	19,445	131	20,232
Total	14,203	13,932	35,198	2,783	66,116

Panel B. Duplications by Racial Groups

Race by number of Duplicates					
# Duplicates	White	Black	Hispanic	Other	Total
0	10,073	9,849	12,189	1,870	33,981
1	1,851	2,005	4,121	340	8,317
2	761	829	2,346	160	4,096
3	437	411	1,724	75	2,647
4	212	210	1,250	56	1,728
5	153	135	1,000	39	1,327
6	120	79	877	26	1,102