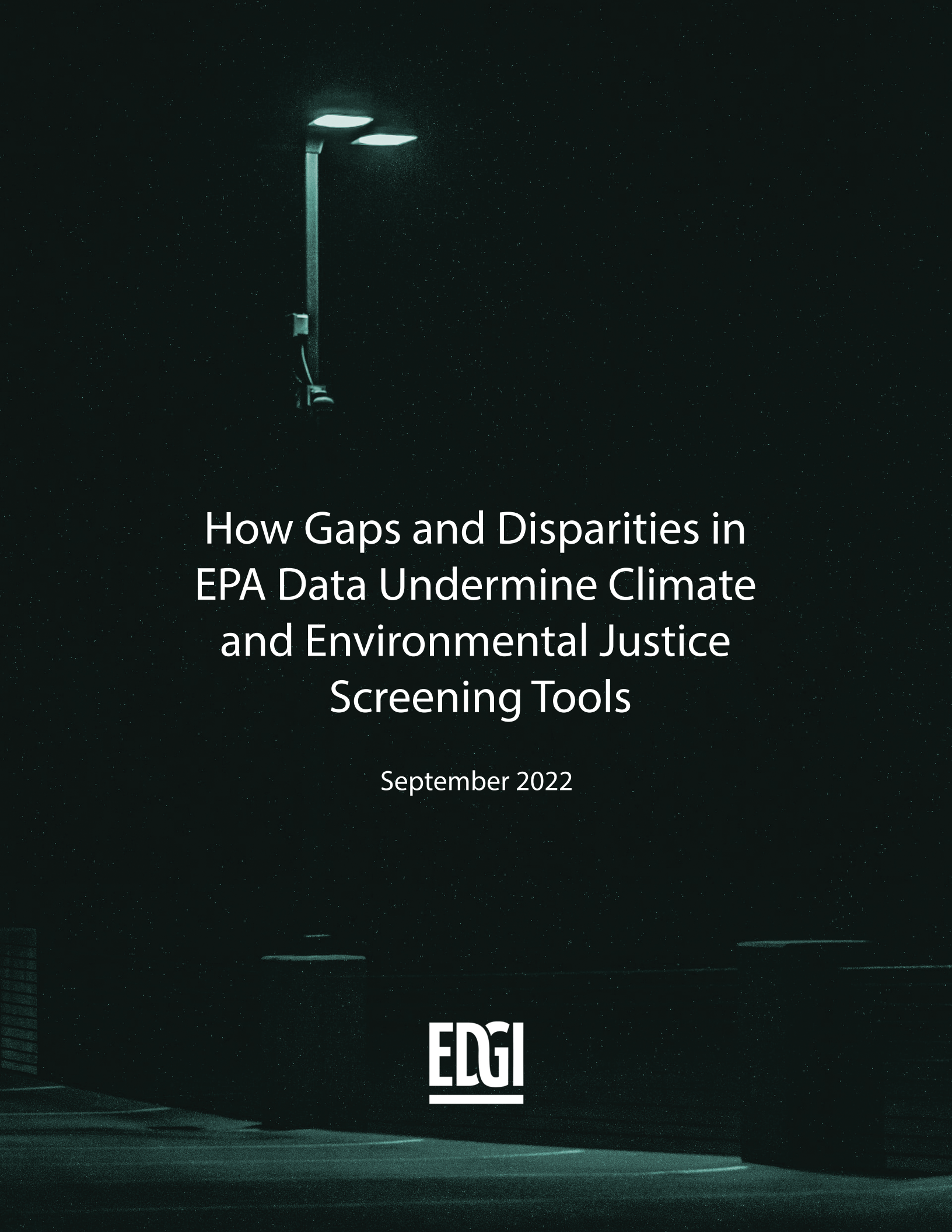


How Gaps and Disparities in EPA Data Undermine Climate and Environmental Justice Screening Tools

September 2022

EDGI

How Gaps and Disparities in EPA Data Undermine Climate and Environmental Justice Screening Tools

September 2022

Authors: Eric Nost, Sara Wylie, Olivia Chang, Olin College PInT, Kelsey Breseman, Steve Hansen, Lourdes Vera, and EDGI

Cover photo: Taofeek Obafemi-Babatunde

The Environmental Data & Governance Initiative (EDGI) is a North American network with members from numerous academic institutions and nonprofit or grassroots organizations, as well as caring and committed volunteers and employees who come from a broad spectrum of work and life backgrounds. EDGI promotes open and accessible government data and information along with evidence-based policymaking.

Summary

Webmaps that are meant to evaluate and "screen" neighborhoods for environmental injustices have seen a lot of interest in both the United States and Canada lately. From informing where to distribute climate funding in the US as "Justice 40" to Canada's Bill C-226, the pursuit of environmental equity has led to a strongly felt need for data and mapping tools that overlay environmental health with racial and income disparities. Indeed, this approach was even enshrined as law in the recent Inflation Reduction Act, which included Rep. Cori Bush and Sen. Ed Markey's Environmental Justice Mapping Bill. But maps made with incomplete or inaccurate data can create false impressions with real-world consequences — like reinforcing existing disparities.

EDGI's research with the Public Interest Technology team from Olin College (PInT) begins to explore gaps and disparities in the data the US Environmental Protection Agency (EPA) uses to track violations and enforce compliance of the Clean Water, Clean Air, Resource Conservation and Recovery, and Safe Drinking Water Acts (CWA, CAA, RCRA, SDWA). We use open and replicable analytical methods to make counts of missing data in these major programs. Key findings include:

1. Over 19,000 facilities regulated under foundational environmental protection laws are missing basic information such as their latitude and longitude. Nearly all — 19,657 out of 19,675 (99.9%) — of these are SDWA-regulated facilities.
2. Data needed for basic EJ assessments, such as the percent minority population surrounding a facility or the Census block it resides in, is missing for 14% of the facilities in EPA's most public-facing database. This increases to 83% of facilities regulated under SDWA.
3. Nationally, the typical facility regulated under each of these environmental protection laws is missing:
    a. 86% of CWA-specific information
    b. 86% of RCRA-specific information
    c. 71% of CAA-specific information
    d. 40% of SDWA-specific information
4. Facilities in majority-minority communities have somewhat worse data quality scores than facilities in majority-white communities, for all acts except SDWA.
5. Data missingness is substantially worse for facilities in areas already screened by EPA to be of particular concern for environmental injustices and majority-minority areas when looking at Clean Water Act inspections in particular.

a. 78% of all facilities regulated under the CWA are missing inspection counts, but only 75% of facilities in majority white areas, rising to 83% of facilities in majority-minority areas.
b. Western states including Texas, New Mexico, Colorado, Utah, and Nevada are much worse when it comes to inspection data completeness for facilities in majority-minority communities.

We interpret these findings in light of what EDGI members have dubbed "environmental data justice," a framework for examining the role of data in environmental justice in terms of "who benefits from data-driven decision-making, what counts as data, and what constitutes data ownership."[1] We believe that to design accurate and just climate and environmental justice screening programs, public agencies must:

1. Conduct and publish analyses of missing data across their programs
2. Investigate systematic reasons for missing data
3. Establish metrics to reflect improvements in data completeness

It's important from a community perspective to not delay action on environmental injustices because of missing data and data disparities, but to add improving data completeness as an overall objective for Justice 40 and similar initiatives.

Datafying Environmental Justice

A large and growing body of research analyzes disparities in how environmental health protection laws like the Clean Water Act are enforced by the EPA and complied with by industry. For instance, three recent papers demonstrate racial and income disparities in Clean Water Act inspections and clustering of SDWA, RCRA, CWA, and CAA violations in specific regions.[2] In identifying environmental injustices, they measure differences in permit violations across racial and income disparities. But if inspections and violations go underreported in a way that correlates with these same disparities, such research can be confounded.

[1] Vera, L. A., D. Walker, M. Murphy, B. Mansfield, L. M. Siad, J. Ogden, and EDGI. 2019. When data justice and environmental justice meet: formulating a response to extractive logic through environmental data justice. Information, Communication & Society 22 (7):1012–1028.

[2] Hui, I., J. Coyle, and A. Ryzhik. 2021. Spatial clustering of hazardous waste, water, air violations in the US. *Environmental Research Letters* 16 (8):084004.

Konisky, D. M., C. Reenock, and S. Conley. 2021. Environmental injustice in Clean Water Act enforcement: racial and income disparities in inspection time. *Environmental Research Letters* 16 (8):084020.

Mueller, J. T., and S. Gasteyer. 2021. The widespread and unjust drinking water and clean water crisis in the United States. *Nature Communications* 12 (1):3544.

At the same time, there is growing interest in both the US and Canada in the development of "screening tools" that allow EPA staff and the public to visualize cumulative impacts on — and justify allocation of funding to — "overburdened" communities. One of the first of these tools — California's CalEnviroScreen — informed the development of EPA's own EJScreen. As part of what the Biden administration calls "Justice 40", the White House Environmental Justice Advisory Council hopes such tools can be extended and developed to ensure that funding from infrastructure and climate bills can be distributed to the communities that have been the most harmed by toxics and fossil fuels. The recently passed Inflation Reduction Act in the US includes provisions requiring investments in environmental justice data and mapping. In Canada, Bill C-226 in Parliament calls for a federal-level database that would identify and track disparate harms from environmental pollutants.

The use of data to more effectively allocate resources to overburdened communities is laudable. However, advancing environmental equity in this way is only possible if the underlying data is accurate and reliable. Unfortunately, we are well aware that it is not!

We set out to illustrate data ambiguities, uncertainties, gaps, and disparities in order to inform the high-level discussions around justice screening tools. Such data gaps matter because if they are used in efforts like Justice 40, they will determine which places count as EJ communities and which don't — and thereby, who does and does not receive needed resources.

Data Gaps

The premise of screening tools is to overlay information about exposures to toxicants — such as water quality information, air quality information, and so on — with socio-economic variables such as income, age, and sometimes race. This is done within and across geographic units like Census tracts in order to assess relative burdens. For example, we can determine the 40% neighborhoods in a city, state, or even the country, most exposed to toxics above a certain indicator of socio-economic vulnerability.

However, if data is missing or incorrect about any of these toxics exposure, socioeconomic vulnerability, or associated geographies, then we are drawing a misleading picture of relative burden.

Researchers have already observed that certain measures of exposure, especially as they relate to water, are not easy to come by.[3] It is somewhat easier to measure how someone or groups of people are exposed to air pollutants than to water contaminants. This presents a data gap, but in the context of EJ screening tools, it is not necessarily critical if the missingness is uniform or, in other words, if all neighborhoods have equally missing data. However, if some neighborhoods have well-measured toxics exposures, but other neighborhoods don't, then the priorities produced by the screening tool could be biased in turn. Specifically, if wealthier and whiter neighborhoods tend to have better monitoring than others, then these places could be weighted higher in the exposure metric than places where exposure is simply not measured.

What We Did

It's possible to measure these kinds of data gaps and disparities. EDGI developed an open source data science tool — a Jupyter Notebook — that determines the scope and distribution of data that screening tools may be missing. Our notebook — which you can view and even run for yourself here — measures how many industrial facilities across the US are missing basic information like location as well as information about things like compliance with environmental protection laws. In plain language, we simply counted the number of blank cells in a spreadsheet. The notebook then assesses whether facilities missing these sorts of measures are more likely to be in majority minority and other socio-economically marginalized areas.

Our analysis centers on one of the most wide-ranging datasets on regulated facilities, the EPA's Enforcement and Compliance History Online (ECHO) database. ECHO isn't currently used in screening tools like EJScreen or the Climate and Economic Justice Screening Tool (CEJST), which leaves out a lot of important information on industry compliance with environmental protection laws and state and federal enforcement of them. But if screening tools such as CEJST did include information from ECHO about specific facilities and their compliance with environmental protection laws, it would carry over significant data gaps that vary — as we will show — by socio-economic measures such as race and linguistic isolation.

[3] Lee, C. A Game Changer in the Making? Lessons From States Advancing Environmental Justice Through Mapping and Cumulative Impact Strategies. https://www.elr.info/sites/default/files/article/2021/07/51.10676.pdf

What We Found

Many facilities lack records about vital information such as location

A few thousand actively regulated facilities that EPA tracks lack even the most basic unit of information: an official "registry" ID with the agency. Thousands more are missing location coordinates. Even when location information is recorded, it is not necessarily correct; it may be an estimate based on a street address, or it may have been incorrectly entered (e.g. flipping latitude and longitude). Almost 7% of facilities with listed latitudes and longitudes have coordinates that are just the center of the county they're in, which is hardly precise.

```
# Facilities missing REGISTRY_ID:
3,530 facilities
0.29 %
-------------------------
# Facilities missing FAC_LONG:
19,674 facilities
1.60 %
-------------------------
# Facilities missing FAC_LAT:
19,674 facilities
1.60 %
-------------------------
# Facilities missing FAC_STATE:
19,808 facilities
1.61 %
-------------------------
# Facilities missing FAC_COUNTY:
123,830 facilities
10.08 %
```

94% of these are SDWA-regulated facilities. It is beyond our ability at this moment to ground truth every record in ECHO, but there are some clear errors. More contextual information — such as the county the facility resides in — is missing at even greater rates. 1 in 10 facilities have no such information available.

```
# Facilities missing FAC_PERCENT_MINORITY:
170,437 facilities
13.87 %
-------------------------
# Facilities missing FAC_DERIVED_CB2010:
174,140 facilities
14.18 %
-------------------------
# Facilities missing EJSCREEN_FLAG_US:
176,930 facilities
14.40 %
```

Missing data of this type is a major barrier to analysis. Missing IDs means that facilities in this database cannot be matched to other databases or referred to programmatically. Missing or incorrect location data prevents many types of geographic analysis: summaries by region, understanding of cumulative effects in an area, and environmental justice analyses that compare areas with certain demographies to others, for example.

And as a result, about 1 in 7 facilities also lack the kinds of records that would help users of ECHO gain insights into the environmental justice dimensions of environmental protection laws like the Clean Air Act. That is, these facilities lack any info in columns such as "EJSCREEN_FLAG_US", a flag to indicate whether a facility is in a Census block group that scores "highly" on EJScreen's indices. These data tend to be missing when geographic coordinates are missing because locating a facility is a necessary first step to recording its context (e.g. the percent of nearby population that is minority). This missingness varies widely between programs. Only 1% of facilities regulated under RCRA are missing environmental justice information. However, 83% of facilities regulated under SDWA lack it.

Many more facilities lack records about their compliance with environmental protection laws…

Typical (Median) Facility in State/Territory is Missing this Percentage of Relevant CWA Program Data					
AK	86	KY	57	OH	57
AL	57	LA	86	OK	57
AR	86	MA	86	OR	86
AS	79	MD	86	PA	86
AZ	86	ME	86	PR	86
CA	86	MI	86	RI	71
CO	86	MN	71	SC	71
CT	86	MO	86	SD	86
DC	86	MP	86	TN	86
DE	57	MS	71	TX	86
FL	86	MT	86	UT	86
GA	86	NC	57	VA	86
GU	86	ND	86	VI	71
HI	86	NE	71	VT	86
IA	57	NH	86	WA	71
ID	86	NJ	86	WI	86
IL	86	NM	86	WV	86
IN	86	NV	86	WY	86
KS	86	NY	71		

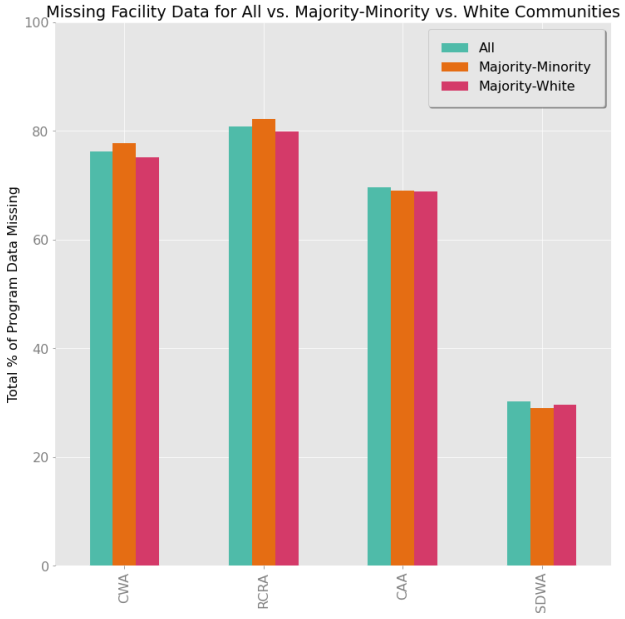
Records of enforcement and compliance measures are even more frequently missing from facility reports. In the Clean Water Act — which is supposed to have the most robust data infrastructure of EPA's programs — most facilities have very little information available about when they were last inspected, how many enforcement actions have been taken, or any penalty amounts levied against them. Nationwide, a typical CWA facility is missing 86% of this information. In other words, the average facility regulated under the CWA's National Pollutant Discharge Elimination System has no entries for 6 out of 7 columns we identified

as relevant for the program.[4] For the CWA, we identified 7 relevant columns facilities should have information for:

- CWA Compliance Status
- Amount of CWA Penalties
- Date of Last CWA Penalty
- CWA Informal Enforcement Actions Taken
- CWA Formal Enforcement Actions Taken
- CWA Inspections
- Days Since Last CWA Inspection

Each of these should have a value for each facility — we should be able to say the facility has had zero inspections, ten penalties, etc. But we find that, typically, facilities have non-null values for only one out of these seven columns. There are good reasons why we see null, or, empty, values: EPA itself may truly not know whether a facility has ever been inspected or in violation and may not want to mislead the public by substituting zero for null. US EPA may not be able to acquire the right information from their state counterparts. Nonetheless, the agency responsible for overseeing the administration of the nation's environmental protection laws should be able to definitely say there's been zero inspections or zero penalties.

This is all information that could otherwise prove useful in a screening tool to understand which neighborhoods are most burdened by limited oversight and routine non-compliance with environmental protection laws. Permit violations are often discovered through inspection; if we cannot reliably audit inspection frequency for an area, we cannot draw reliable conclusions about permit non-compliance. Similarly, if penalty amount information is missing, we cannot know if enforcement is equitably applied.

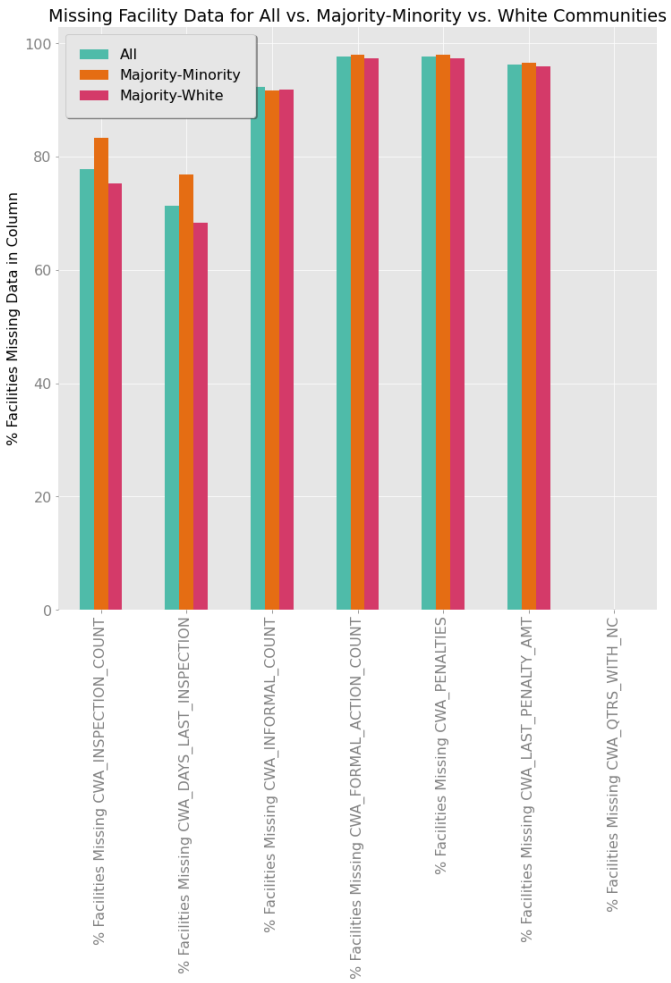


[4] Although they lack information on environmental justice, SDWA-regulated facilities tend to have more complete information about their compliance with that act - the typical SDWA facility is missing "just" 40% relevant records.

...and these gaps in data vary spatially and socio-economically

Facilities surrounded by majority minority neighborhoods (> 50% nonwhite) have somewhat less information about inspections and other variables than facilities in majority white areas (see chart). For instance, CWA-regulated facilities in majority minority neighborhoods are, in total, missing 77% of relevant CWA information; CWA-regulated facilities in majority white areas are missing 75%. A similar trend holds true for areas flagged by EJScreen as potential areas of concern for environmental injustices; CWA-regulated facilities in these areas are missing 3% more data than facilities. As the chart shows, racialized differences in data completeness are not always huge and vary by program.

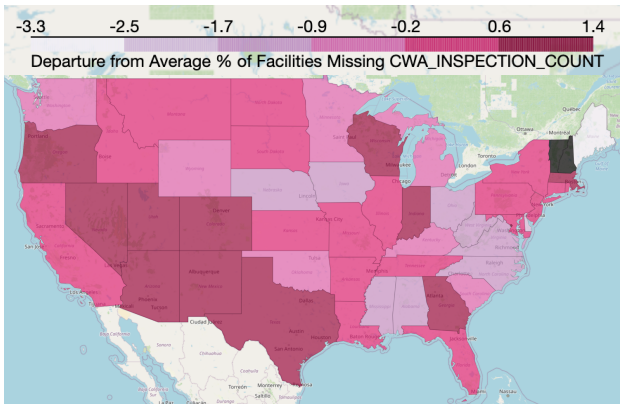
But these differences are especially clear when we look at CWA inspections. 83% of facilities in majority minority Census blocks are missing information on the number of inspections they've faced over the past 3-5 years, compared to just 75% of facilities in white neighborhoods missing this information. The CWA is supposed to have the most complete set of records — more complete than CAA, RCRA, or SDWA — suggesting that the disparities in those programs might be more severe if the data were more complete. Because permit violations are often found through inspections, this suggests likely underreporting of violations specifically in minority communities.



The largest disparity in information we see in the existing EJ screening tool CEJST is that CWA facilities in communities with high levels of "linguistic isolation" — measured as the "Percent of households where no one over the age of 14 speaks English well" (whatever that means) — have 3% less information about the compliance of facilities in their area. For other measures that CEJST tracks, facilities in "better off" neighborhoods — those

with more high school graduates and higher incomes — actually have less complete information than facilities in other areas. This may be related to how CEJST measures education and income rates at the Census tract level rather than the more circumscribed Census block level.

Disparities in data quality also vary by state. In some states, majority minority neighborhoods are more likely to have better information about facilities near them. Though some commentators argue screening tools are ready to roll out anywhere, this may not actually be the case because of such state-specific data inequalities. As shown in the map, majority minority communities in states like Maine or Iowa have somewhat more complete information on CWA inspections than majority minority communities nationwide. Such communities in states like Kentucky and Pennsylvania are about on par. But majority minority communities in states like Texas, Indiana, and Wisconsin are home to a greater percentage of facilities with no CWA inspection information.



Conclusions

Data on industry non-compliance and governments' lack of enforcement of environmental protection laws exists, but isn't being utilized in screening tools, putting communities in harm's way. At the same time, much of this data is riddled with gaps and inaccuracies. While some observers are hopeful that any state could adopt EJ screening tools, missing data and data errors often vary by state, especially in the federal EPA's datasets.

Missing data in particular is a crucial challenge because it can be misinterpreted to mean a facility is operating legally or that a community has avoided harm — "There are no records of violations here? We should turn our attention elsewhere to communities that do have recorded violations." In fact, one common data cleaning practice is to simply remove rows with missing data, which in this case might specifically discount some of the most harmed communities. Instead, the most cautious approach to missing data might assume worst-case scenarios: that facilities with missing data do have violations, for example. But not all columns with missing data have a maximum to which they could be set — penalty amount, for example. And assuming worst-case scenarios in areas with missing data could

again overlook real harms in favor of presumptive harms. In short, data missingness makes data-based screening tools inherently unjust.

There are many reasons elements of enforcement and compliance such as violations may not even be measured. For instance, less funding turns into fewer inspections and, ultimately, violations can go unobserved. As such, even if the federal government could somehow close the data gaps we have identified by ensuring all regulated facilities have complete records, we would not necessarily be solving underlying problems where politics and lack of funding stymie robust auditing programs.

For this, we need systems that also ensure data is properly understood and contextualized. Accurate, just climate and environmental justice screening programs would require agencies to:

1. Conduct and publish analyses of missing data across their programs
2. Investigate systematic reasons for missing data
3. Establish metrics to reflect improvements in data completeness

There is an inherent tension here, however. On the one hand, we have suggested that data gaps and disparities may prove harmful in climate and environmental justice screening tools, and that more data collection is necessary (putting environmental governance on a kind of "data treadmill"). On the other hand, the injustices wrought by toxics and fossil fuel industries demand urgent redressing — while doing so with incomplete data may reinforce inequalities. These tensions are ones we believe the framework of environmental data justice is meant to wrestle with. We believe, following the precautionary principle, that a lack of data is no excuse not to move forward with opportunities to invest in communities.