ENVIRONMENTAL RESEARCH
LETTERS

LETTER • OPEN ACCESS

Spatial clustering of hazardous waste, water, air violations in the US

ENVIRONMENTAL RESEARCH
LETTERS

LETTER

Spatial clustering of hazardous waste, water, air violations in the US

Iris Hui*, John Coyle and Abraham Ryzhik

Bill Lane Center for the American West, Stanford University, Room 172, Y2E2, Stanford, CA 94305, United States of America
* Author to whom any correspondence should be addressed.

E-mail: irishui@stanford.edu

Keywords: air quality violations, water quality violations, hazardous waste violations, Environmental Protection Agency (EPA), spatial analyses

OPEN ACCESS

RECEIVED
21 December 2020

REVISED
26 June 2021

ACCEPTED FOR PUBLICATION
1 July 2021

PUBLISHED
20 July 2021

Abstract

We examine spatial patterns of three types of Environmental Protection Agency violation, hazardous waste, water, and air quality, at the facility level. Since facilities operate independently, our null hypothesis is that their violations should be spatially randomly distributed. That is, we do not expect to observe spatial clusters of violations. In addition, systemic factors such as socio-demographic characteristics as well as environmental justice indices should not correlate with violations. Empirically, we find both hypotheses are refuted. Our findings confirm that environmental inequalities have been exacerbated by underlying social vulnerabilities, particularly in the case of Native Indian territories, which consistently show a disproportionately high number of environmental violations. We identify 'hot spots', spatial clusters where the number of violations is higher than expected in such a way that cannot be explained by socio-demographic or environmental factors. These hot spots call for local case studies to further investigate causes of spatial clustering of violations.

1. Introduction

Access to clean water, clean air and sanitation is recognized by the United Nations as a basic human right. In the United States, the Environmental Protection Agency (EPA) is the primary independent government agency responsible for guaranteeing this right by monitoring the relevant aspects of environmental health throughout the country. Established in December 1970 under President Richard Nixon, the agency has since played a vital role in conducting environmental assessment, research, outreach activities, as well as ensuring environmental justice.

The EPA defines environmental justice as '[t]he fair treatment and meaningful involvement of all people regardless of race, color, national origin, or income with respect to the development, implementation, and enforcement of environmental laws, regulations, and policies'[1]. In turn, fair treatment is defined by the axiom that 'no population, due to policy or economic disempowerment, is forced to bear a disproportionate share of the negative human health or environmental impacts of pollution or environmental consequences resulting from industrial, municipal, and commercial operations or the execution of federal, state, local and tribal programs and policies' (Brulle and Pellow 2006). This is often referred to in the environmental justice literature as 'distributive justice', where distributive principles are designed to cover the distribution of benefits and burdens of economic activities among individuals in a society (Olsaretti 2018).

Numerous studies have documented a correlation between social vulnerabilities, environmental hazards and environmental injustice. Social vulnerabilities are typically defined as 'the susceptibility of a given population to harm from exposure to a hazard, directly affecting its ability to prepare for, respond to, and recover [from it]' (Cutter and Emrich 2006, Cutter *et al* 2009). Cutter *et al* (2003) expand this notion and argue that social vulnerability is 'partially the product

[1] www.epa.gov/environmentaljustice

of social inequalities–those social factors that influence or shape the susceptibility of various groups to harm and that also govern their ability to respond.' Social vulnerabilities, when interacting with place-based inequalities (such as the condition of the built environment and level of economic activities) and place-based environmental hazards tend to result in environmental injustice, where disadvantaged communities often shoulder a larger share of the environmental burden.

One important early finding was a 1983 U.S. General Accounting Office study examining the communities surrounding the four major hazardous waste landfills in the South. All of the communities were found to be disproportionately African American. Another seminal work was a 1987 study which found that race was the most important factor in predicting where toxic waste facilities would be located (Chavis and Lee 1987). Mohai *et al* (2009) observed unequal distribution of toxic and hazardous waste facilities with respect to minority and low-income communities, and Maantay *et al* (2010) conducted a meta-analysis and documented how social vulnerabilities, together with proximity to hazards would lead to adverse health outcomes and disproportionate impacts on communities with high vulnerabilities.

Studies have identified an array of vulnerability factors such as socio-economics (e.g. race, income, gender), age, housing conditions, isolation (such as language barriers, lack of transportation option) (Li *et al* 2010; Boer *et al* 1997, Morello-Frosch *et al* 2001, Konisky 2009). The major factors that influence social vulnerability are believed to be 'lack of access to resources (including information, knowledge, and technology); limited access to political power and representation; social capital, including social networks and connections; beliefs and customs; building stock and age; frail and physically limited individuals; and type and density of infrastructure and lifelines' (Cutter *et al* 2003, 2008). These pre-existing social vulnerability and environmental injustice factors have been linked to a variety of outcomes, including a higher rate of heart-related illnesses or death, poorer levels of air quality and higher chances of air pollutant related illnesses, and higher rates of infectious diseases. (See Cooley *et al* 2012 for a review summary of this literature.) These social vulnerability factors are not unique to the U.S.: other cross-national studies reveal a similar pattern of social vulnerability (Fekete 2009).

The goal of this paper is to investigate the extent to which environmental justice is upheld through an examination of spatial heterogeneity and clustering of EPA violations at over a million facilities in the US over a period of three years, between 2016 and 2019. There are two layers of environmental justice. The first layer is the systemic factors, such as socio-demographics, that contributed to a higher level of violation. The second layer is additional, excessive spatial concentration of violations. We explore how much spatial heterogeneity and clustering can be explained with recourse to conventional social vulnerability factors already identified in the existing literature, and how much of the variation in excess may be due to unmeasured factors, such as local institutional characteristics or inspection and enforcement practices.

In this paper, the term EPA violation indicates both non-compliance with environmental law(s) discovered in an onsite civil inspection or procedure as well as criminal investigations into such matters as the illegal disposal of hazardous waste, the illegal discharge of pollutants into a water source, the improper removal and disposal of asbestos materials, etc. While facilities are not randomly located but rather tend to cluster together, our null hypothesis expects each to operate independently from its neighbors. Hence, we posit under the null hypothesis that these violations would be spatially-randomly distributed, a notion which is understood to include two components. First, we would expect no significant spatial pattern of violations across existing facilities; that is, one firm's violation should not be correlated with its neighboring firms' violations. Secondly, we would expect an absence of correlation between systemic factors and spatial clustering. That is, we would not expect socio-demographic characteristics such as poverty level or racial composition to be correlated with the number of violations. In other words, no specific areas or types of communities should be overburdened by environmental violations.

2. Materials and methods

2.1. Data

We created our dataset by webscraping[2] the EPA ECHO ('Enforcement and Compliance History Online') website which archives the monitoring and enforcement actions of all registered facilities in the US. We focused on violations in three areas: hazardous waste, water, and air quality[3]. We obtained data on violations, socio-demographic characteristics, and environmental justice indices[4] from the website. Since the environmental justice indices are highly correlated, we used factor analysis to summarize the variables into one Environmental Justice index, which we refer to as 'pollution burden'. More information on these variables can be found in the appendix, table A1.

[2] Our webscraping methodology can be found at https://abrahamryzhik.github.io/BLCSummer19/29a7f5eb267929b9578af261b074cd509c1ab7f8/index.html

[3] ECHO also keeps data on drinking water. However, because demographic variables and location information (i.e. longitude and latitude) are absent in most of the records, we cannot include drinking water in our analyses. We accessed the dataset between 9 and 14 July 2019. https://echo.epa.gov/

[4] EJ Screen https://echo.epa.gov/help/facility-search/water-search-results-help

All variables obtained from the EPA ECHO site are interpolated values created by the EPA and represent the demographic composition of the area immediately surrounding each facility within a three-mile radius. This in turn educes several limitations. First, the socio-demographic data were obtained from the 2010 Census which pre-dated our study period. Second, the racial/ethnicity composition came from two separate Census Tables. The racial composition variable consists of the categories, namely, White (alone), Black/African American, American Indian and Alaska Native, Asian, Native Hawaiian and Other Pacific Islander, and two or more races. The Hispanic population, on the other hand, is only measured represented by a binary variable (yes or no). In other words, the variable representing White or Black percentages does not differentiate between Hispanic and non-Hispanic Whites or Blacks.

In addition, the EPA provides two seemingly similar, yet substantially different variables. One is 'Indian Country', a binary variable which denotes if a facility is located geographically within a Native American territory. The other is a continuous variable which measures the percentage of Native Americans within a three mile radius. As numerous Native Americans have settled outside of the Native American territories, the two variables are not collinearly correlated and hence do not pose a multicollinearity problem to our equation.

It is important to note that the data we scraped did not contain information on the specific number, nor the type, of violations for each facility. Instead, the data contained a column, 'quarters with violations', which shows the number of quarters (maximum 12) within a three year period that the facility was in violation of any environmental statute.

We focused on violations in three areas: hazardous waste, water, and air quality. Given that some of the contextual factors are not available for Hawaii and Alaska, we restricted our analysis to the 48 continental states and Washington D.C. Altogether we have about 325 000 (about 300 000 completed cases); 560 000 (430 000); and 179 000 (137 000) facilities in the water, hazardous waste, and air datasets respectively.

2.2. Statistical methods

We first plotted the spatial patterns of violations. Drawing insight from the existing environmental justice literature, we employed a hierarchical linear model (HLM) with random state effect to analyze the correlation between systemic factors and the number of quarters in violation. Individual facilities are our smallest unit of analysis. The equation we ran is:

$$Y_i = \alpha_{ji} + \beta X_i + \varepsilon_i$$

where the dependent variable, Y, is the number of violations within the past 12 quarters for each facility (i). This variable ranges from 0 to a maximum of 12. X contains all socio-economic characteristics and our measure of pollution burden. We also included the industry codes (i.e. North American Industry Classification System, or 'NAICS' codes) as dummy variables in the model to account for variation in violation tendency across industrial composition. Lastly, each state (j) has its own varying intercept. Online Appendix A1 lists and explains the variables in the model.

We obtained the residuals from the HLM model. These residuals represent the portion of variation that cannot be explained by the independent variables which we included (i.e. socio-economic characteristics, EJ summary index, industry types and state characteristics). We then explored spatial dependency in these residuals. We created a spatial weight matrix where spatial neighbors are defined as other facilities within a one-mile radius. We used Local Indicator of Spatial Association (LISA) to identify spatial 'hotspots', where there are unexplainable clusters of violations (Anselin 1995). Online Appendix A2 and A3 explains our choice of the weight matrix and spatial models respectively.

3. Results

3.1. Spatial variation of violations

We began by plotting the number of quarters in violation at the facility level. Figures 1(a)–(c) show the distribution of EPA violations for water, air and hazardous waste. The dependent variable ranges from 0 to a maximum of 12 quarters of violations in the past three years. The darker dots indicate a higher number of violations.

Comparing these three figures reveals that there are more facilities with quarters in violations for water facilities than for air or hazardous waste. Several dark color clusters may be observed in figure 1(a), notably in Washington-California corridor, Midwest, South and the east coast. There are relatively fewer dark color clusters in figure 1(b) (air quality) or figure 1(c) (hazardous waste), the exception being the Central Valley in California which is known for its air pollution.

3.2. Correlations with socio-demographics and Environmental Justice index

The expectation of our aforementioned null hypothesis would be that no statistical relationship should exist between socio-economic characteristics and EJ index. Table 1 shows the results of our HLM model where numerous statistically significant relationships are observed. Among them, the dummy variable 'Indian Territory' (which indicates whether a facility is located in Native American territory) is consistently positive and statistically significant in all three models. This finding shows that facilities in Native American territories exhibit a higher number of quarters

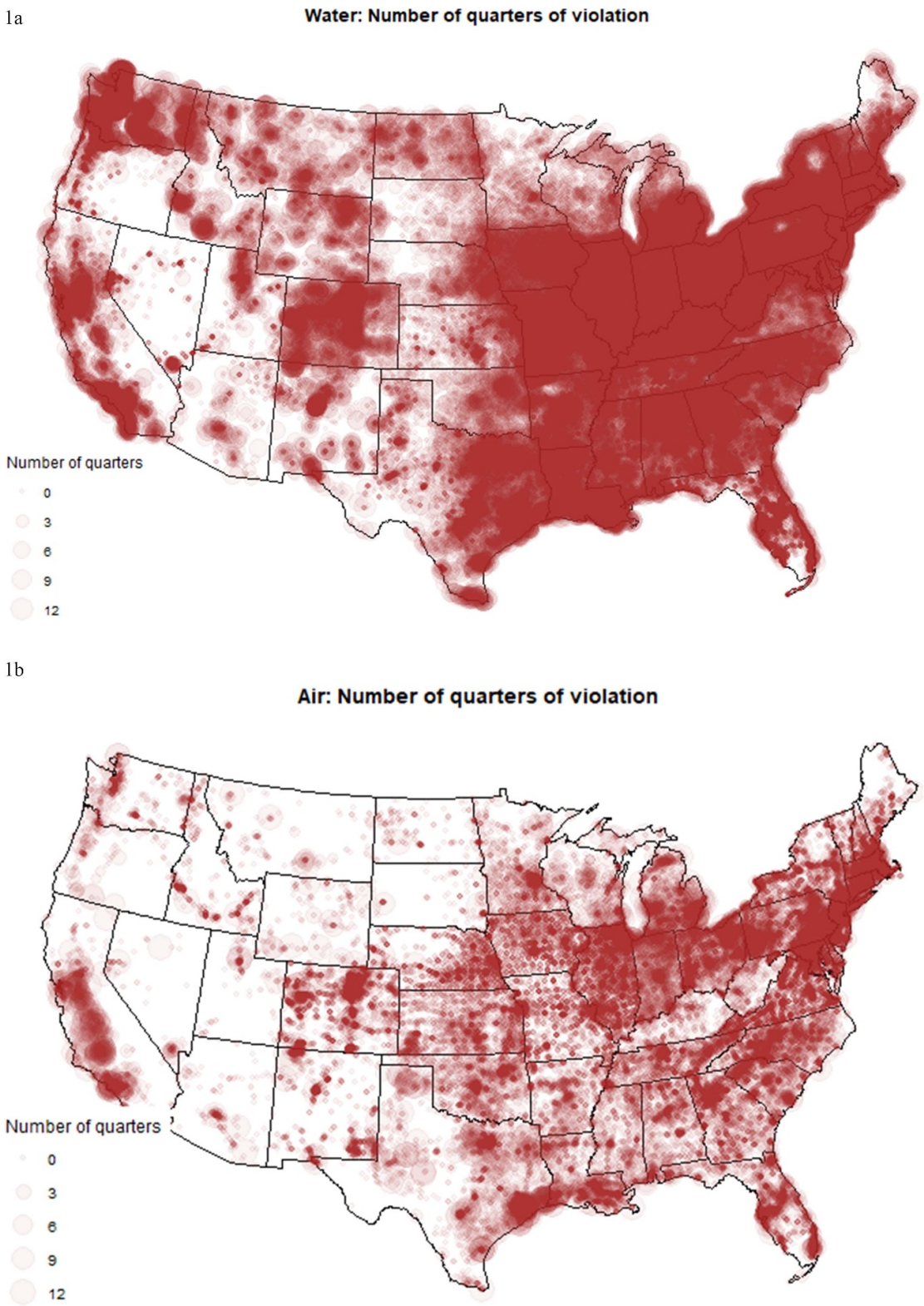


Figure 1. Spatial distribution of number of quarters with EPA violations for water, air and hazardous waste datasets. Note: Figures plot the number of quarters in violation in the past 12 quarters.

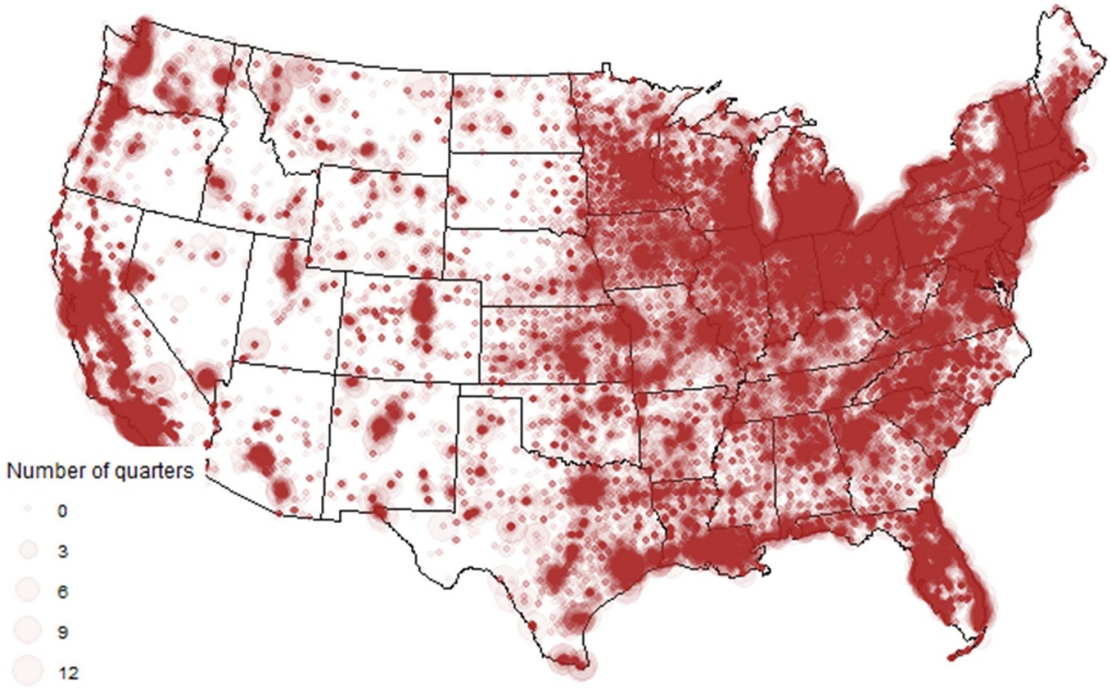


Figure 1. (Continued.)

Table 1. Results from HLMl.

	Water	Air	Hazardous waste
(Intercept)	1.458***	0.126**	0.085***
	(0.193)	(0.048)	(0.024)
Population density	−0.000***	0.000	−0.000**
	(0.000)	(0.000)	(0.000)
% households on public assistance	−0.312	0.202	−0.286**
	(0.253)	(0.182)	(0.092)
% households under poverty line	0.398***	0.057	0.079*
	(0.068)	(0.043)	(0.032)
Indian Country	1.820***	0.127*	0.319***
	(0.108)	(0.053)	(0.046)
% Black	−0.003***	0.000	0.000
	(0.000)	(0.000)	(0.000)
% Hispanic	−0.005***	0.000	−0.000
	(0.000)	(0.000)	(0.000)
% Asian	−0.000	−0.003**	0.001*
	(0.001)	(0.001)	(0.000)
% Native American	−0.011***	0.001	0.002*
	(0.001)	(0.001)	(0.001)
% Senior	0.011***	−0.000	−0.001*
	(0.001)	(0.001)	(0.000)
% BA degree or above	−0.011***	−0.002***	−0.001***
	(0.001)	(0.000)	(0.000)
% Income more than $75k	0.001	0.002***	0.001
	(0.001)	(0.000)	(0.000)
Pollution burden	0.026***	−0.012***	−0.004
	(0.006)	(0.004)	(0.002)
NAICS Categories	Included	Included	Included
Number of states	49	49	49

Note: we ran three separate HLM models where facilities are nested within states. *** $p \leqslant 0.001$; ** $p \leqslant 0.01$; * $p \leqslant 0.05$.

in violation of air, water, and hazardous waste statutes than those outside these territories. In contrast, the average level of education, measured in percentage of population with a bachelor's degree or above, is consistently negatively correlated with quarters in violation.

There are other causes for concern. One indicator of economic prosperity, % households under poverty line, is positively correlated with the number of violations in the water and hazardous waste dataset. Similarly, the coefficient for % senior population is positive (0.011), indicating a slight disadvantage for areas with higher concentration of elderly people. Turning to the pollution burden indicator, we find a positive correlation with the number of violations in the water dataset but a negative one in the air dataset.

Furthermore we find, contrary to expectation, that facilities in areas with a higher white population, on average, have the highest level of water violations. This finding, which contradicts the results one would expect having read the environmental justice literature, is, we believe, partly a result of the way that the race/ethnicity variables were constructed by the Census and EPA[5], and partly driven by the fact that a lot of violations occurred in the rust belt and former industrial heavy areas with a high concentration of poor Whites.

In the hazardous waste dataset, we observed that both Indian Country variable and the % Native American are both positive and statistically significant. That is, facilities in Indian Country areas, as well as areas with a higher concentration of Native Americans were worse off than those with a higher cluster of Whites alone.

3.3. Unexplained spatial clusters with excessive concentration of violations

As discussed in the Methodology section, we also applied a LISA model on the residuals from the HLM models and identified 'hot spots' (i.e. clusters that are 'high-high' and statistically significant at 0.05 level). First, our hotspot analysis identified facilities that have residual values from our HLM model that were higher than expected. Then, we geolocated these facilities to see if there was a pattern of spatial clustering. Figures 2(a)–(c) display the remaining spatial clusters that cannot be explained by the variables in table 1. These hotspots point to potential areas with spatial concentrations of excessive environmental burden.

Overall, these hotspot cases usually account for less than 3% of the total number of facilities in our datasets. For the three datasets, we identified about 8200 hotspots for water (i.e. 2.9% of all cases), 2500 for hazardous waste (0.6% of cases) and 580 for air quality (0.4% of cases). As shown in figure 2(a), on the West Coast, three major clusters in Seattle in Washington, and in the Bay Area and the Los Angeles metro in California, clearly stand out. In the Mountain West region, the Denver metro in Colorado has more clusters than in neighboring states. The Houston and Dallas metros are two major clusters in Texas. The pattern varies substantially in the Midwest and east coast. Missouri, for example, has numerous clusters spread across the state that are not confined to any specific metro area. Similarly, Louisiana exhibits a spread-out pattern like Colorado. In addition, hotspots are concentrated across the rural rust belt (that spans from Syracuse/Rochester, NY to Indiana, Ohio and Illinois) and the coal/mining industrial rust belt corridor (that runs from Pennsylvania to Kentucky)[6].

Using the same methodology, we found fewer hotspots in the air and hazardous waste datasets. In figure 2(b), the only cluster that stands out on the West Coast is the Central Valley in California. On the East Coast, Pennsylvania, a state with heavy industrial facilities, likewise exhibits a heavy concentration of hotspots. Several metros also stand out in this figure such as Houston, Texas, Miami, Florida, Detroit, Michigan, Chicago, Illinois. In figure 2(c), in contrast to figure 2(b), the two clusters in California are in the Bay Area, Los Angeles, and the San Diego metro again, instead of the Central Valley. The rust belt consistently has been plagued by hotspots in all three figures.

4. Discussion

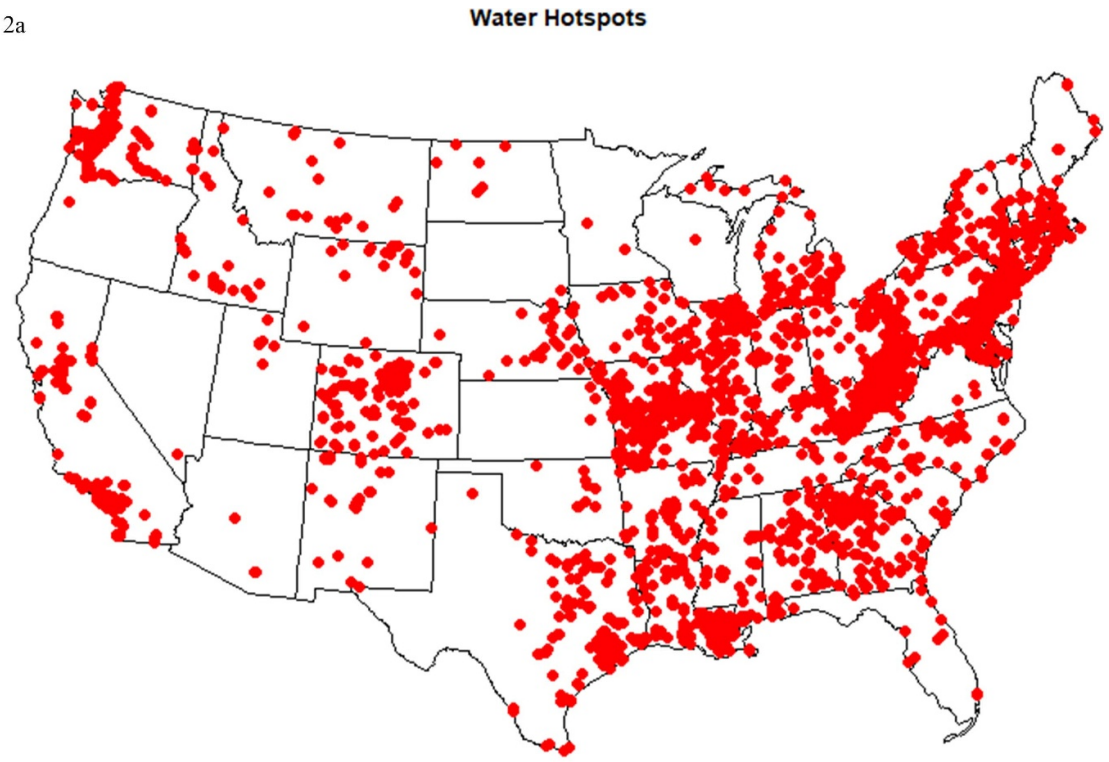
Our study is both pertinent and critical for two reasons. First, recent decades have witnessed a slow divestment by the EPA of its central administrative power, and a deferral of many of its duties to state and local governments. Under the Obama administration, the EPA had an average of at least 17 000 full-time employees, a number which steadily declined during the Trump presidency so that in 2019, the EPA only employed around 14 000 staff full-time[7]. This reduction comes at a time when, according to some reports, many key monitoring activities have devolved to state EPAs. In 2018, the EPA issued a memorandum explicitly giving up much of its authority to state EPAs (USEPA 2018). Such increases in fragmentation come at a time when the Environmental Data and Governance Initiative, in a 93-page report on the EPA under the Trump administration, identifies a 'plummet in enforcement' and a 'retreat' by the EPA from its central role as a steward of American public health (Fredrickson 2018). This pattern of fragmentation in turn creates the possibility for irregularities in environmental law enforcement across communities. Mapping out heterogeneity in

[5] See discussion above on how the variables, e.g. White, contain both White Hispanic and White Non-Hispanic in one.

[6] See the geographic definition of 'rust belt'. https://beltmag.com/mapping-rust-belt/

[7] www.epa.gov/planandbudget/budget

2a



2b

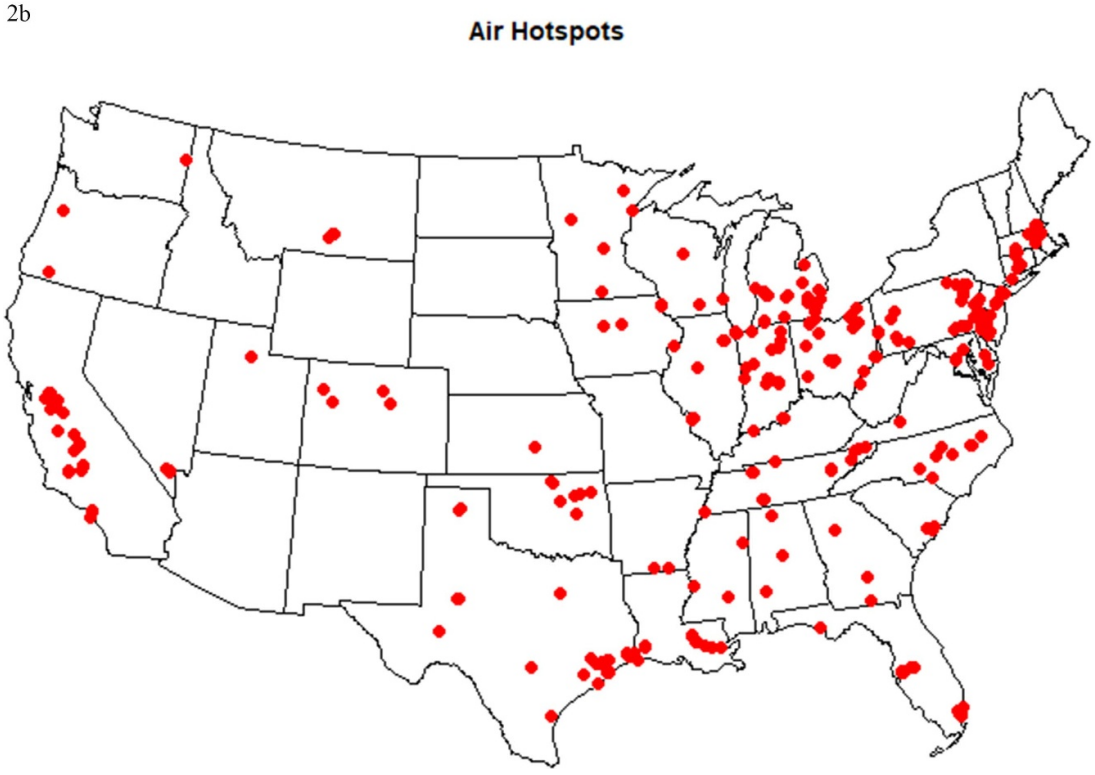


Figure 2. Spatial distribution of hotspots for water, air and hazardous waste datasets.
Note: figures plot the hotspots (i.e. spatial cluster of facilities with higher than expected number of violations) that remained unexplained by our HLM models.

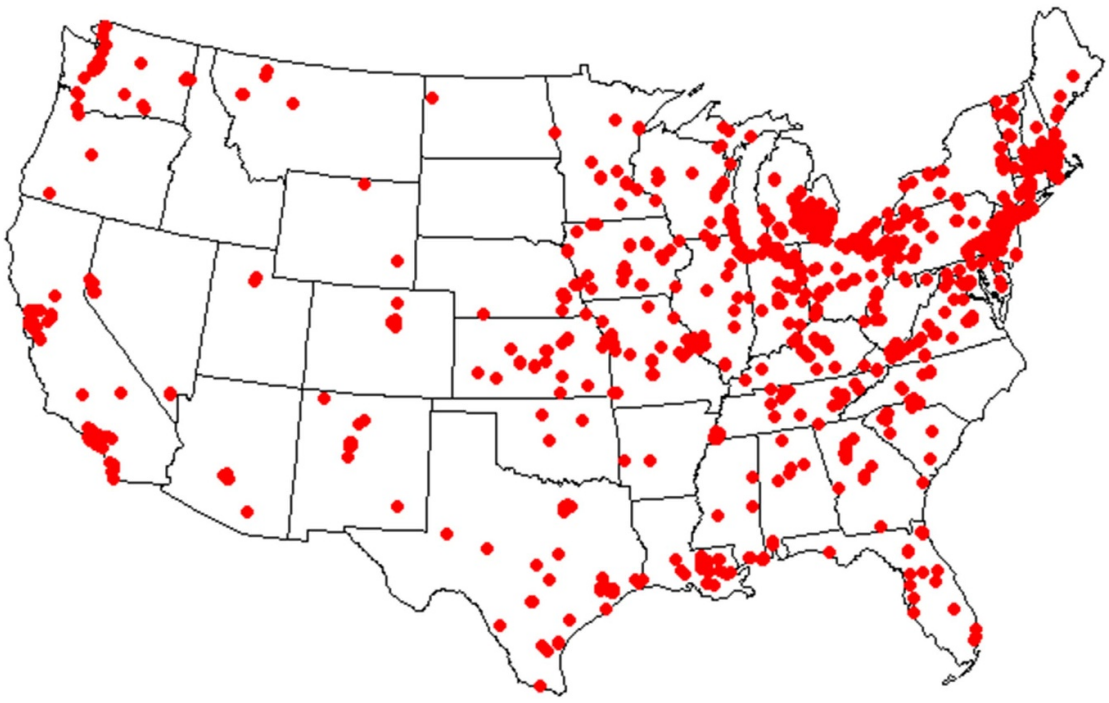


Figure 2. (Continued.)

the number of violations across states and within states helps to identify the resulting discrepancies, if any, in how local authorities handle their monitoring responsibilities. Secondly, our hotspot analysis identifies areas with unexplained clustering which in turn provides a roadmap for further studies.

Our study has several notable limitations. First, we focus primarily on identifying patterns instead of explaining why these patterns emerged. To accomplish the latter, at least two additional projects are necessary. The first would be an examination of variation across states to explain heterogeneity in EPA violations. Since we have controlled for North American Industry Classification System (NAICS) categories in the HLM models, differences in industrial composition can be ruled out. One possible explanation may have to do with variation in each state's regulatory capacity. To measure that, one would have to examine factors such as the number of state-EPA staff, budgets, regularity, and attitudes toward monitoring activities, etc. According to previous studies, another key to understanding these patterns may lie in the regulatory bodies themselves—specifically, whether officials on the local boards are appointed or elected (Mullin 2009). Boards composed of elected representatives are often found to have more vigorous monitoring practices than boards that consist primarily of government officials (Hoover 2008).

A second limitation of our study is that our dependent variable, quarters in violation, does not capture nuances such as the number of violations within each quarter or their severity. Minor violations such as missing a routine report are substantially different from major violations such as illegal dumping of waste or water discharge, a subtlety which this dataset does not recognize. In addition, we do not have information about who reported these violations—that is, whether the violations were spotted during regular inspections or through citizens' complaints, information which is crucial for identifying the role of private citizens' monitoring of compliance with environmental standards and laws.

A third limitation is that the racial composition variables (e.g. White, Black) obtained from EPA are not separated from ethnicity. Ideally, we would prefer to have separate variables for Non-Hispanic White and Hispanic Black or Non-Hispanic White and Hispanic White etc, as these would allow us to isolate the effects of race and ethnicity. We suspect the mixing of race and ethnicity is one of the primary reasons why we observed a higher level of violation in communities with a higher concentration of Whites.

When analyzing the spatial patterns of violations, two possible, yet contradictory hypotheses may account for their existence. First is a lack of stringent oversight, where facilities may disregard environmental procedures and laws. According to this hypothesis, spatial patterns would be seen in states where EPA enforcement is lax. A second hypothesis would posit that clusters would arise in states where oversight is more stringent since these states are more

likely to identify violations. This would lead to the presence of clustering in states with stricter enforcement of standards and laws. Because we do not have data on how local EPAs function, nor their efficacy in monitoring facilities within their jurisdictions, we cannot at this time provide a full explanation for why some areas or states tend to perform better or worse than the national averages.

Another consideration is that in this paper, we focus primarily on our dependent variable, 'quarters of violations'. One should note here that the amount a facility is in violation does not necessarily correlate with the amount such a facility pollutes. Many of the violations are likely related to missing paperwork in which case a higher level of violation would not denote a higher level of pollution. Further studies are needed to explore the specific types of violations, which are not available in our datasets. Furthermore, an area with low level of violation may have other forms of environmental injustice occur simultaneously, such as lack of access to clean drinking water, or poor air quality in impoverished neighborhoods.

It is also important to emphasize that the absence of a violation cluster does not necessarily indicate an area does not run afoul of environmental laws. As discussed, the discovery of EPA violations is subject to the honesty and integrity of the relevant enforcement body and institutional neglect may lead to violations in socially vulnerable areas being underreported. This could be an explanation for why we do not always observe a strong correlation between lower income and hotspots, as the literature on social vulnerability suggests. Furthermore, we observe evidence of variables associated with less socially vulnerable communities—higher education (BA degree or above) or high wealth (income >$75k)—being correlated with hot-spots. This finding could be attributed to more rigorous enforcement practices in wealthier communities due to political pressure from the residents. Again, further investigations and case studies are necessary to draw conclusions.

Lastly, our study calls for a follow-up project to conduct localized case studies of the hotspots. We certainly do not believe there is a single, universal, systemic missing variable that explains all the spatial clusters in the nation. Every local cluster requires its own qualitative study of, for example, facility owners/operators, local government officials, or EPA staff. Through interviews, perhaps some insights could be reached concerning what and how local factors play a role.

5. Conclusion

Our HLM results confirm the unfortunate reality that some socio-demographic characteristics are correlated with a higher occurrence of environmental violations, and, therefore, potentially higher risks of exposure to environment hazards. Our hotspot analyses illustrate that the conventional socio-demographics, industry types and environmental characteristics cannot fully explain away why some areas have higher clustering of violations than others. Our hot-spot analyses serve as a starting point for more follow-up studies to explore potential institutional factors that gave rise to these spatial clusters.

Data availability statement

The data generated and/or analyzed during the current study are not publicly available for legal/ethical reasons but are available from the corresponding author on reasonable request.

Appendix

A1. Independent variables used in the HLM model

EPA created these datasets with the methodology detailed in this page. (https://echo.epa.gov/resources/echo-data/about-the-data).

The socio-demographic data came from the 2010 Census block group database, Population and Housing Summary Tape Files 1A and 3A. The environmental indicators came from the Environmental Justice Screening and Mapping Tool, 2010.

We observed a high positive correlation among our EJ variables. To reduce the problem of multicollinearity in regression, we first reduced the ten variables into a single dimension using factor analysis. We extracted one factor, namely, 'pollution burden' which accounted for about 80%–85% of variance of our variables in the three datasets.

Table A1. Explanation of independent variables in HLM model.

Independent variables	Explanation
Population density	Population per square mile
% households on public assistance	Number of households on public assistance divided by total number of households
% households under poverty line	Number of households under poverty line divided by total number of households
Indian Territory	A binary variable indicates inside Indian territory or not
% Hispanic	Percentage Hispanic population
% Black	Percentage Black population
% Asian	Percentage Asian population
% Native American	Percentage Native American population
% Senior	Percentage senior over 65 year old
% BA degree or above	Percentage of population with at least a bachelor's degree
% Income more than $75k	Percentage of population with household income more than $75k
Pollution burden	Summary index of pollution measures. See table A2
NAICS codes	Dummy variables for industry types according to North American Industry Classification System
State codes	All 48 continental states and D.C.

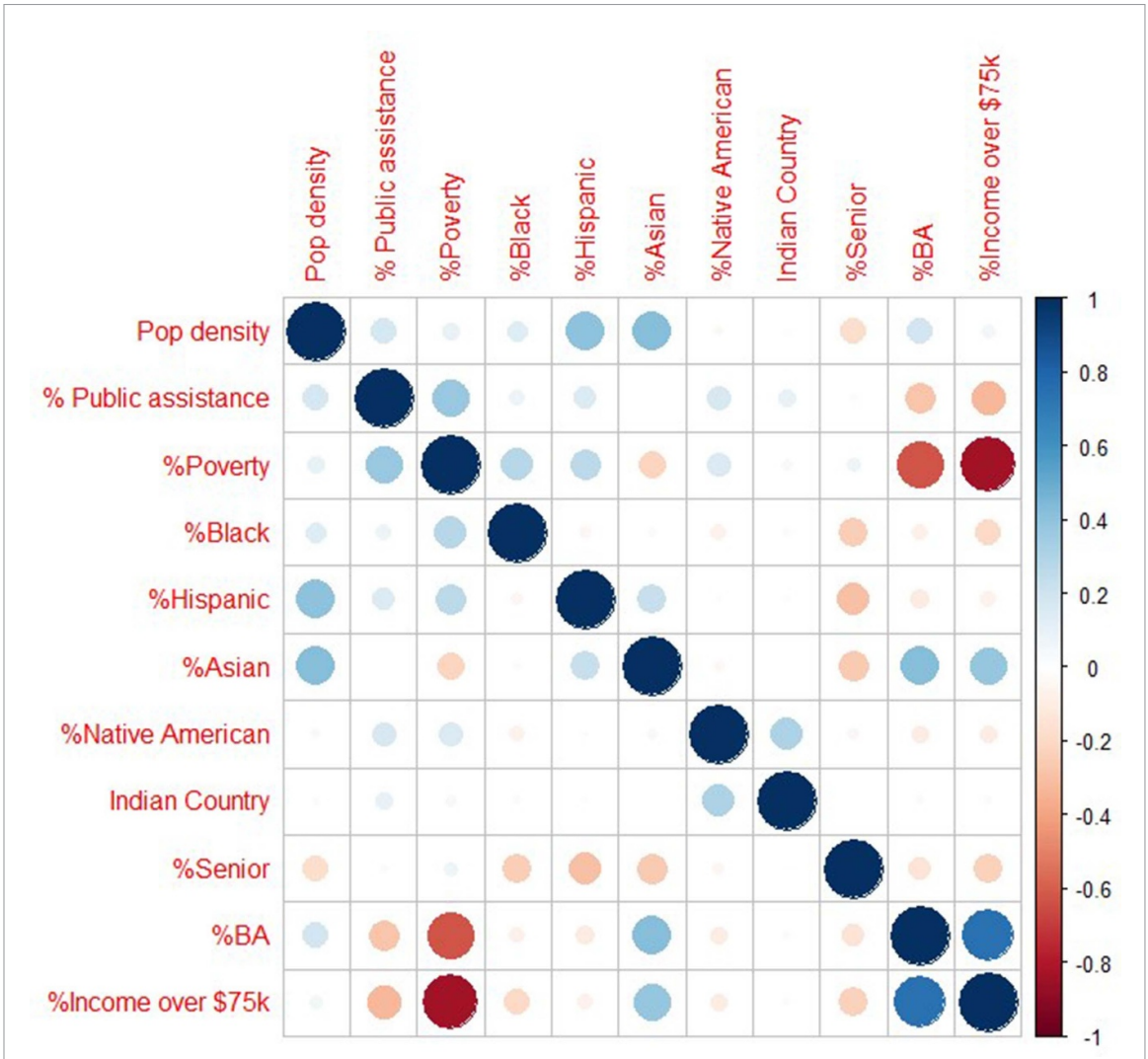


Figure A1. Correlation among socio-economic variables in table A1.

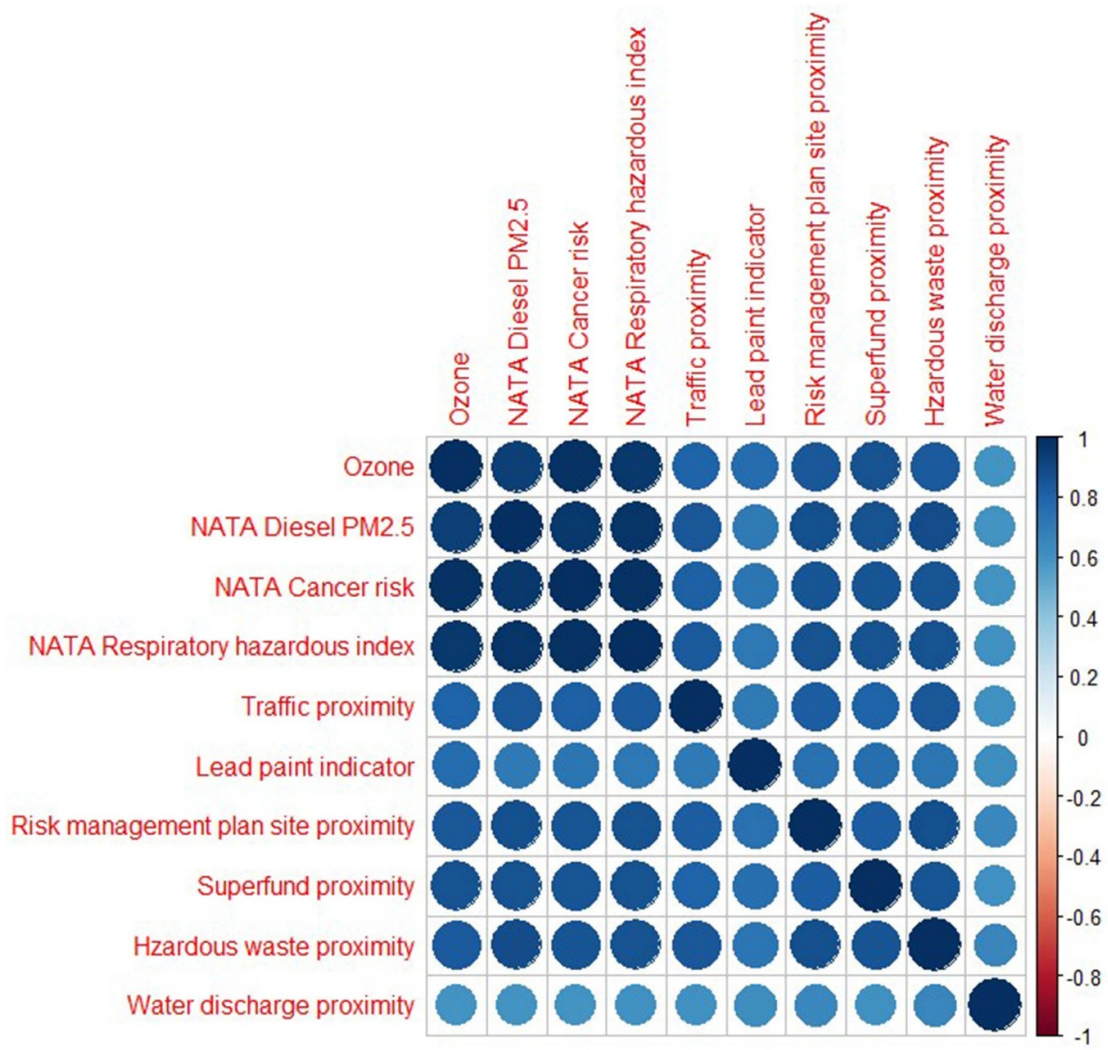


Figure A2. Correlation among all the Environmental Justice variables in the EPA ECHO data.

A2. Creation of spatial weight matrix

For our spatial analyses, we created a spatial weight matrix based on nearest neighbors within 1 mile. We experimented with values between 0.5 mile to 1 mile. When we limited the radius to 0.5 mile, on average, each observation would have only one neighbor. This low number of neighbors rendered spatial analyses useless and irrelevant, as the results would not differ from those of a non-spatial model. When we extended the radius to 1 mile, each observation, on average, had 14 neighbors. We believe one mile, which represents a typical walking distance, offers a balance between over- and under-accounting for spatial autocorrelation.

Figure A4 shows the spatial correlogram which shows the extent of spatial autocorrelation in the water dataset. The average Moran's I value is about 0.2 within one lag of neighbors. The spatial autocorrelation dropped notably after the first lag.

Table A2. Explanation of independent variables used to create the EJ Summary Index, namely, 'Pollution Burden'.

Independent variables[a]	Explanation
Ozone	Ozone summer seasonal avg. of daily maximum 8-hour concentration in air in parts per billion
NATA Diesel PM2.5	Diesel particulate matter level in air, μg m^{-3}
NATA Cancer Risk	Lifetime cancer risk from inhalation of air toxics
NATA Respiratory Hazard Index	Air toxics respiratory hazard index (ratio of exposure concentration to health-based reference concentration)
Traffic Proximity	Count of vehicles (AADT, avg. annual daily traffic) at major roads within 500 m, divided by distance in meters (not km)
Lead Paint Indicator	Percent of housing units built pre-1960, as indicator of potential lead paint exposure
Risk Management Plan Site Proximity	Count of RMP (potential chemical accident management plan) facilities within 5 km (or nearest one beyond 5 km), each divided by distance in kilometers
Superfund Proximity	Count of proposed or listed NPL—also known as superfund—sites within 5 km (or nearest one beyond 5 km), each divided by distance in kilometers
Hazard Waste Proximity	Count of hazardous waste facilities (TSDFs and LQGs) within 5 km (or nearest beyond 5 km), each divided by distance in kilometers
Water Discharge Proximity	RSEI modeled Toxic Concentrations at stream segments within 500 m, divided by distance in kilometers (km)

[a] www.epa.gov/ejscreen/overview-environmental-indicators-ejscreen.

Table A3. Results from spatial Durbin linear model with spatially lagged right-hand side variables ('lmSLX').

	Estimate	St. error	*p*-value
Population density	0.000	0.000	0.005
% households on public assistance	−0.369	0.381	0.332
% households under poverty line	0.429	0.078	0.000
Indian Territory	1.837	0.194	0.000
% Black	−0.003	0.000	0.000
% Asian	−0.004	0.001	0.000
% Native American	0.000	0.001	0.829
% Senior	−0.011	0.001	0.000
% BA degree or above	0.010	0.001	0.000
% Income more than $75k	−0.010	0.001	0.000
Pollution burden	0.002	0.001	0.003
Population density	0.017	0.008	0.037

Note: included dummy variables for states and.

Results from spatial error model ('errorsarlm').

	Estimate	St. error	*p*-value
Population density	0.000	0.000	0.000
% households on public assistance	−0.284	0.263	0.279
% households under poverty line	0.366	0.073	0.000
Indian Territory	1.806	0.110	0.000
% Black	−0.003	0.000	0.000
% Asian	−0.006	0.000	0.000
% Native American	−0.001	0.001	0.575
% Senior	−0.011	0.001	0.000
% BA degree or above	0.011	0.001	0.000
% Income more than $75k	−0.011	0.001	0.000
Pollution burden	0.001	0.001	0.501
Population density	0.027	0.007	0.000

Note: included dummy variables for states and NAICS.

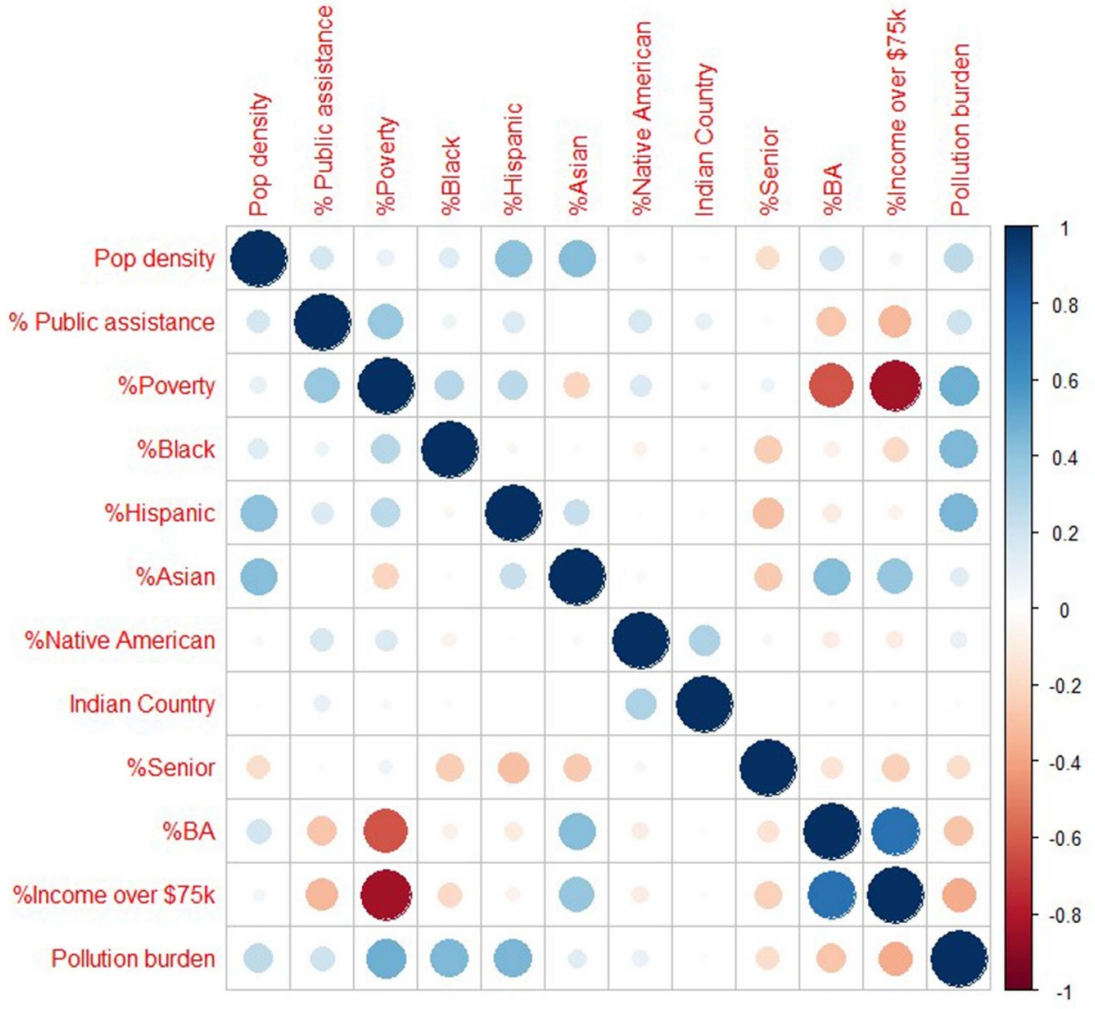


Figure A3. correlation among all the independent variables in the final HLM model.

A3. Spatial regressions

Ideally, we would want to run a hierarchical spatial autoregressive model using the R package 'HSAR'. In order to do that, we would need to convert our spatial weight matrix (which is a 'listw' object) into a 'nb2mat' object. However, R ran out of memory with this process (despite the fact that our machine has 64 GB ram). As a result of these excessive memory demands, we did not conduct the hierarchical spatial autoregressive model.

However, we were still able to run a spatial Durbin linear model with spatially lagged right-hand side variables and a spatial error model, both using the R package 'spatialreg' on our water dataset. While the coefficients are not exactly identical to the ones presented in table A3, both the lmSLX and errorsarlm models generated estimates in the same direction as those found in table A3.

A4. State random effects

We explored whether there was systemic variation across states. Figure A5(a) shows a dot-chart that ranks the magnitude of the state random effects from the HLM model. Since we allow each state to have a varying intercept, these varying intercepts represent the average number of violations for our 48 states plus D.C. area.

Figure A5(a) shows that, on average, Washington state has the highest average number of violations (close to five quarters in violation out of twelve) in the water dataset. Iowa and Michigan rank second and third. Ten states have an average number of quarters in violation higher than 1.

Figures A5(b) and (c) show that, on average, the number of quarters in violation for hazardous waste and air are much lower than that for water. The highest magnitude for the random state effect is less

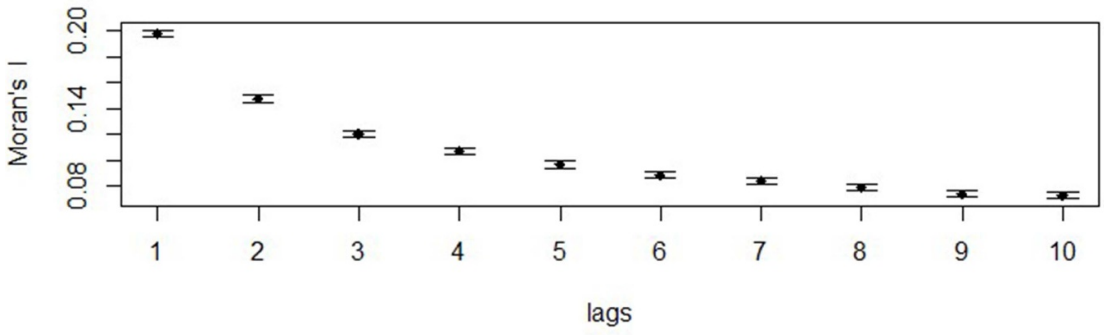


Figure A4. Moran's I value by number of spatial lags

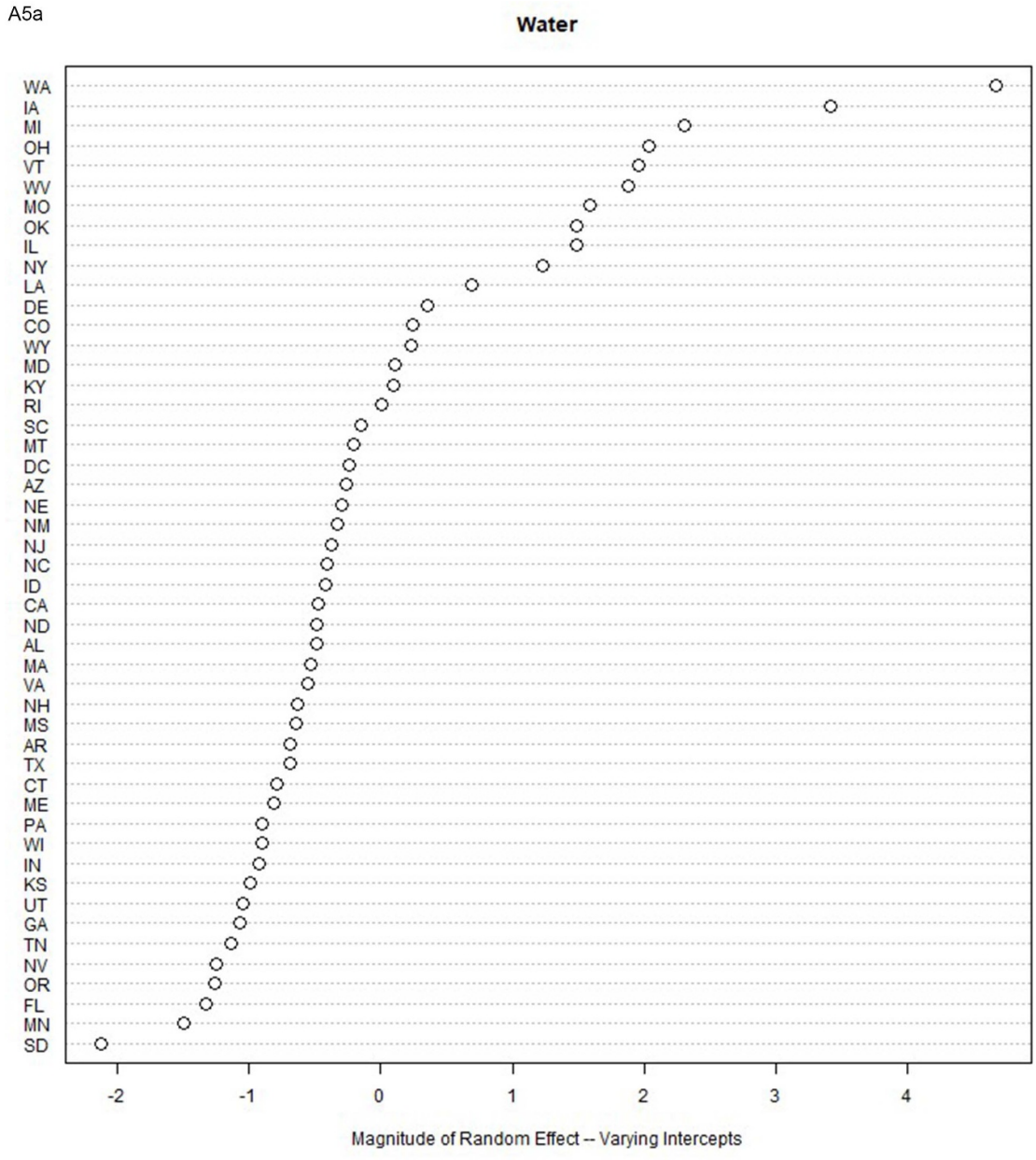


Figure A5. Ranking of states in term of average number of EPA violations for water, air and hazardous waste datasets. Note: We ran three separate HLM models with random state effects. Figures plot the magnitude of random effect. Note that, on average, states have the highest number of violations in the water dataset.

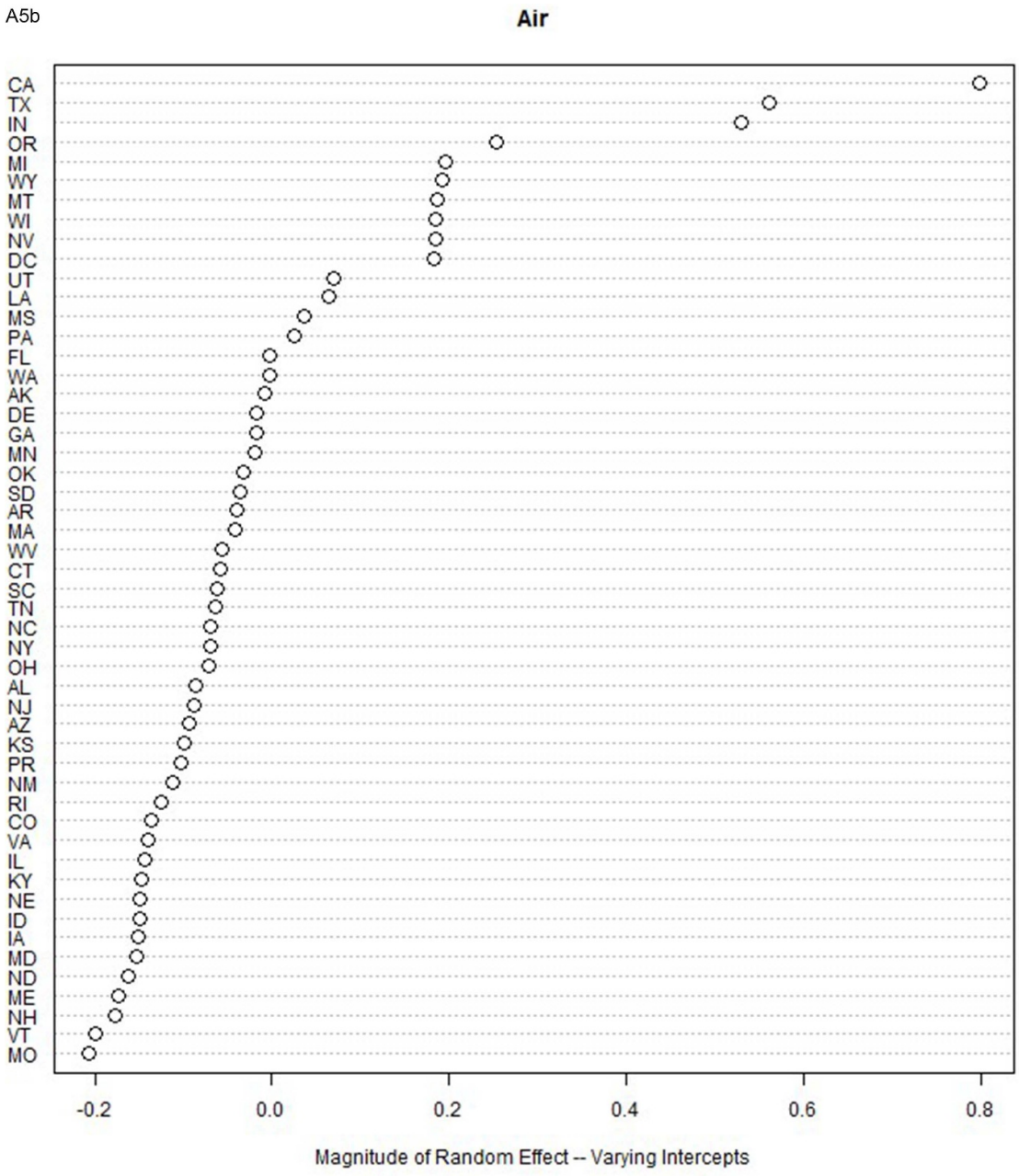


Figure A5. (Continued.)

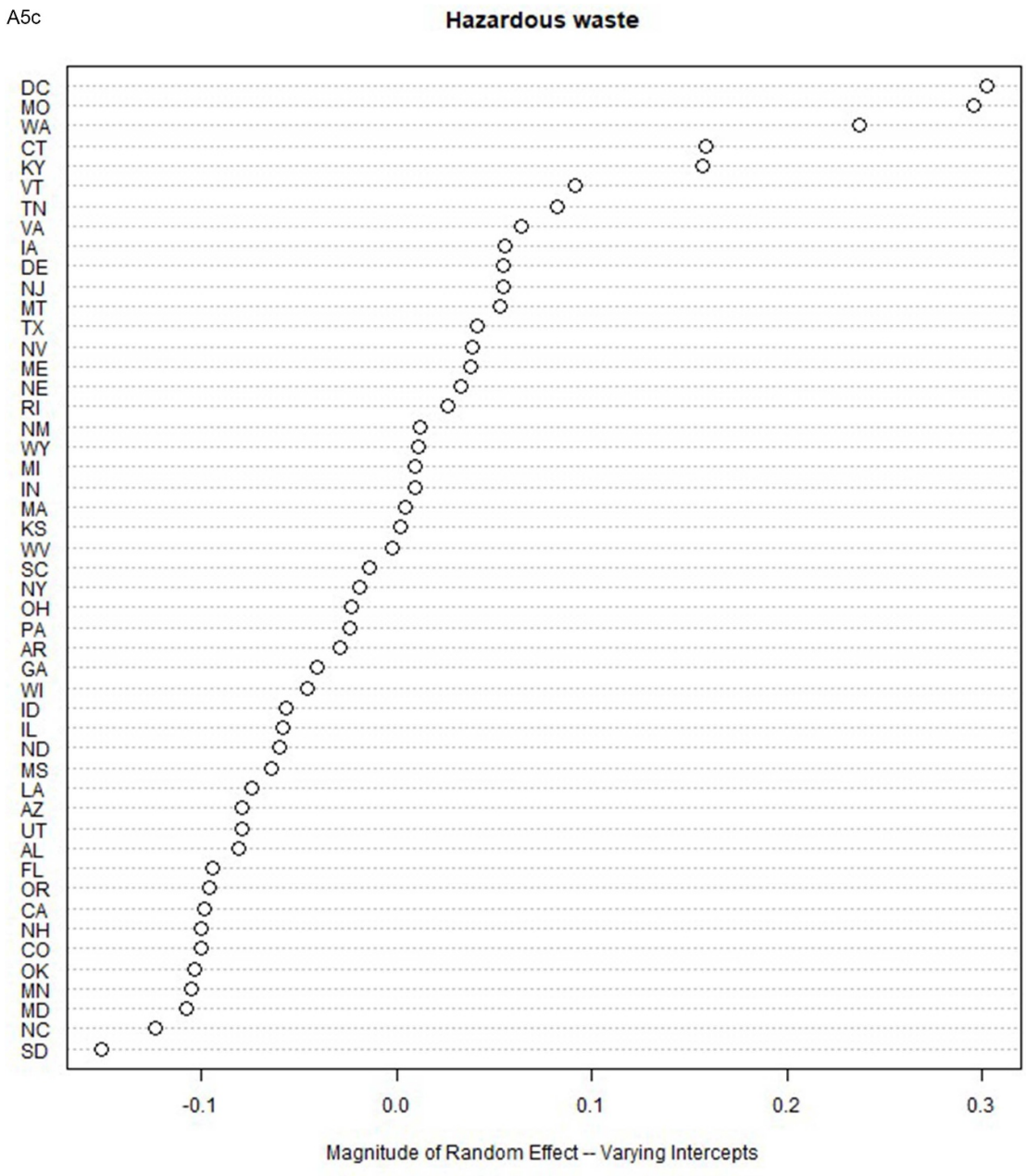


Figure A5. (Continued.)

than 1 quarter in violation out of 12 in the air dataset and less than 0.5 in the hazardous waste dataset. It should be noted that the rankings of states vary substantially in our three Figures, which suggests that no one state bears a disproportionately heavy violation burden in all three domains.

A5. Comparison of hotspots and non-hotspots

We analyzed whether these hotspot facilities are substantially different from those that are not designated as hotspots and exhibit the comparisons in figures A6(a)–(c). To interpret the Figures, note that the red color represents the hotspot facilities and black/grey color denote that of non-hotspot facilities. We plotted the density distribution of our compositional characteristics (e.g. % black population) in the background, and added the means as vertical lines. Take figure A6(a), for example, the first panel shows the distribution of the black population among the hotspot and non-hotspot facilities. The red vertical line is to the *left* of the black vertical line, indicating the mean percentage of black population in the hotspot facilities is lower than that among non-hotspot facilities. In the next panel, as the red and black vertical lines almost overlap, there is a smaller difference in the composition of Latino population between our two samples in the water dataset. We used *t*-test to test the differences in means. All differences are statistically significant at 0.05 level due to our large sample sizes.

In terms of socio-demographic composition, hotspot facilities in the water dataset are more likely to be found in areas with a lower black population, as well as areas with lower fraction of the population with a bachelor's degree or above and with income more than $75k (figure A6(a)). The contrasts are more obvious in the air quality dataset. Hotspot facilities are more likely to be in areas with more minorities, lower fraction of population with bachelor's degree, and income over $75k (figure A6(b)), indicating potential over-burden in socially vulnerable areas. The differences between the hotspot and non-hotspot facilities in the hazardous waste dataset are substantively small with hotspots tending to occur in areas with higher fraction of Black and poor.

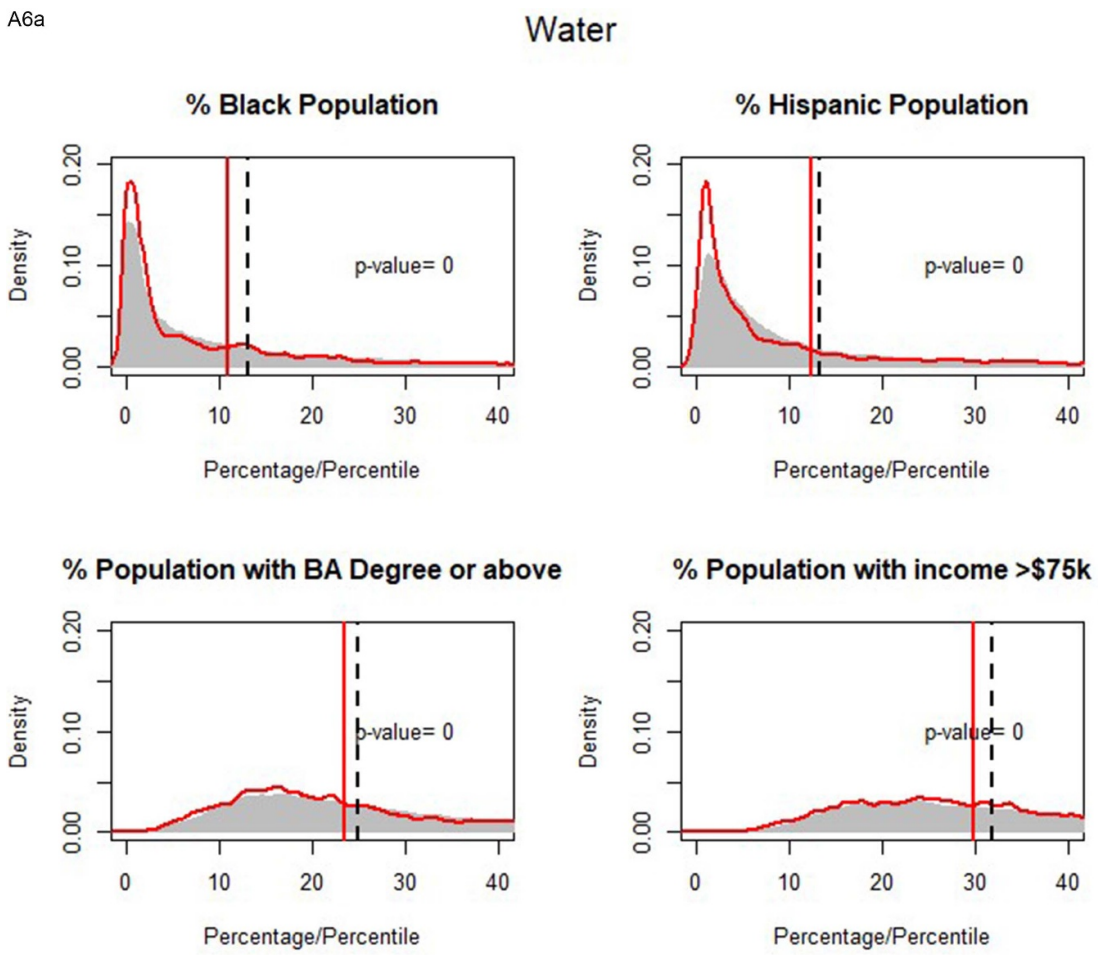


Figure A6. comparison of background characteristics of hotspot vs. non-hotspot facilities.
Note: Note that the red color represents the hotspot facilities and black/grey color denote that of non-hotspot facilities. We plotted the density distribution of our compositional characteristics (e.g. % black population) in the background, and added the means as vertical lines.

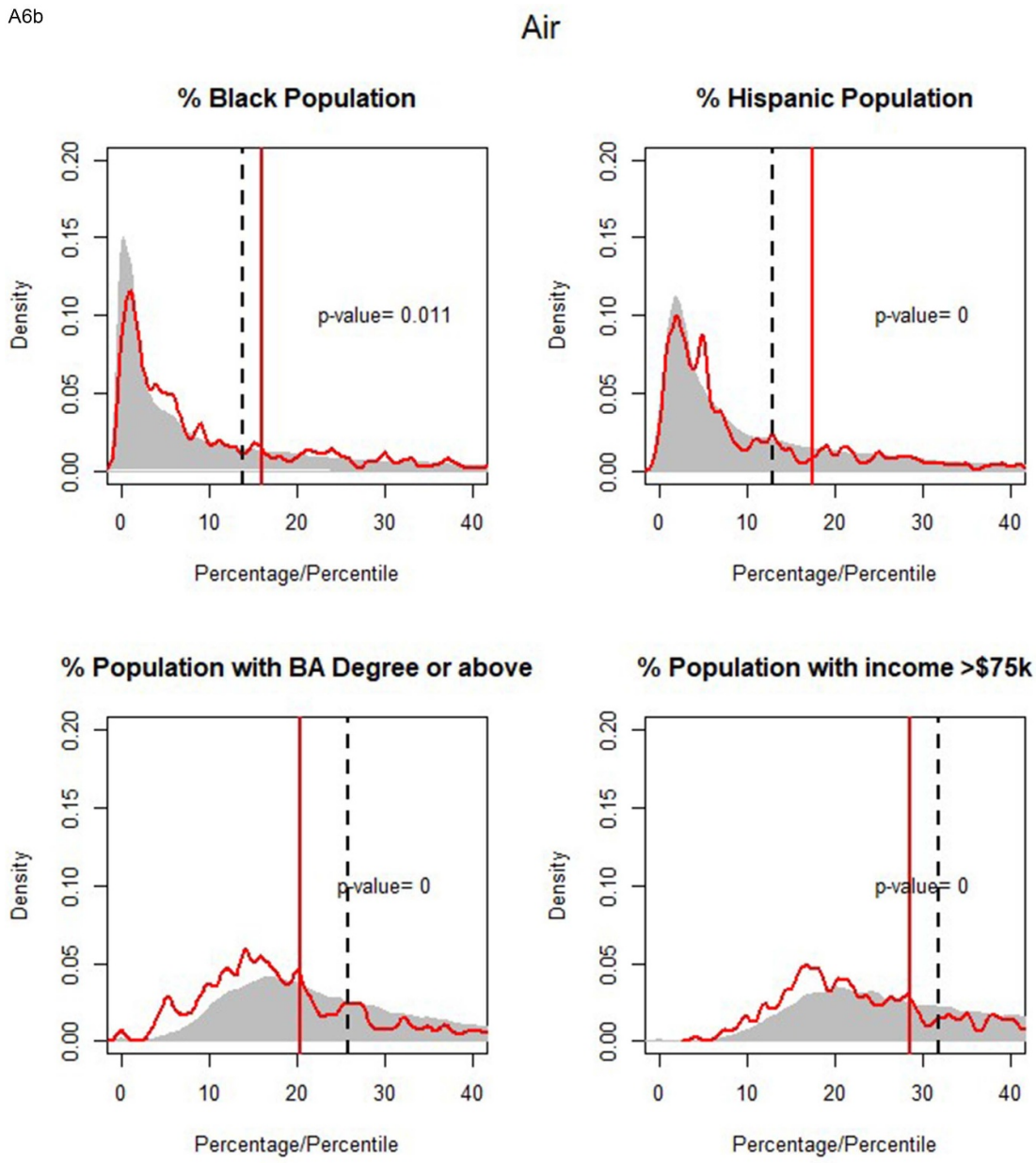


Figure A6. (Continued.)

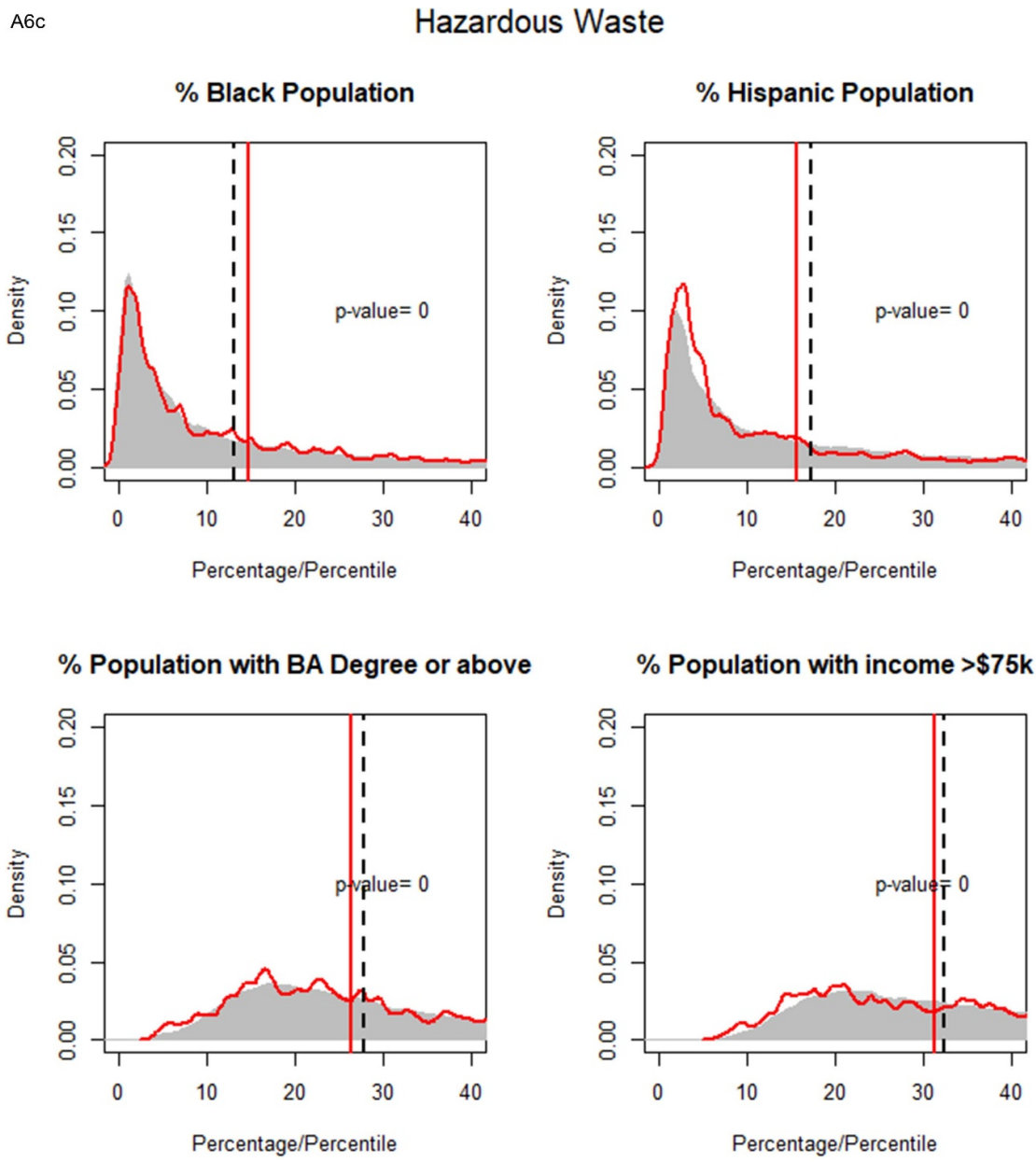


Figure A6. (Continued.)

ORCID iD

Iris Hui https://orcid.org/0000-0003-3351-7957

References

Anselin L 1995 Local indicators of spatial association—LISA *Geogr. Anal.* **27** 93–115

Boer J T, Pastor M, Sadd J L, Snyder L D, Boer J T and Washington G 1997 Is there environmental racism? The demographics of hazardous waste in Los Angeles county *Soc. Sci. Q.* **78** 793–810

Brulle R J and Pellow D N 2006 Environmental Justice: human health and environmental inequalities *Annu. Rev. Public Health* **27** 103–24

Chavis B F and Lee C 1987 *Toxic Wastes and Race in the United States* (New York: United Church Christ)

Cooley H *et al* 2012 Social vulnerability to climate change in California California Energy Commission Publication Number: CEC-500-2012-013

Cutter S L, *et al* 2009 Social vulnerability to climate variability hazards: a review of the literature Final Report to Oxfam America (available at: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.458.7614&rep=rep1&type=pdf)

Cutter S L, Barnes L, Berry M, Burton C, Evans E, Tate E and Webb J 2008 A place-based model for understanding community resilience to natural disasters *Glob. Environ. Change* **18** 598–606

Cutter S L, Boruff B J and Shirley W L 2003 Social vulnerability to environmental hazards *Soc. Sci. Q.* **84** 242–61

Cutter S L and Emrich C T 2006 Moral hazard, social catastrophe: the changing face of vulnerability along the Hurricane coasts *Ann. Am. Acad. Pol. Soc. Sci.* **604** 102–12

Fekete A 2009 Validation of a social vulnerability index in context to river-floods in Germany *Nat. Hazards Earth Syst. Sci.* **9** 393–403

Fredrickson L 2018 updated 2019 A sheep in the closet: the erosion of enforcement at the EPA. Environmental Data & Governance Initiative (available at: https://envirodatagov.org/publication/a-sheep-in-the-closet-the-erosion-of-enforcement-at-the-epa/)

Hoover G A 2008 Elected versus appointed school district officials: is there a difference in student outcomes? *Public Finance Rev.* **36** 635–64

Konisky D M 2009 Inequities in enforcement? Environmental justice and government performance *J. Policy Anal. Manage.* **28** 102–21

Li Z, Konisky D M, Zirogiannis N and Kahn M 2010 Racial, ethnic, and income disparities in air pollution: a study of excess emissions in Texas *PLoS One* **14** e0220696

Maantay J, Chakraborty J and Brender J 2010 Proximity to environmental hazards: Environmental Justice and adverse health outcomes. Prepared for the U.S. Environmental Protection Agency *Strengthening Environmental Justice Research and Decision Making: A Symp. the Science of Disproportionate Environmental Health Impacts*

Mohai P, Pellow D and Roberts J T 2009 Environmental Justice *Annu. Rev. Environ. Resour.* **34** 405–30

Morello-Frosch R, Pastor M and Sadd J 2001 Environmental Justice and Southern California's 'Riskscape': the distribution of air toxics exposures and health risks among diverse communities *Urban Aff. Rev.* **36** 551–78

Mullin M 2009 *Governing the Tap: Special District Governance and the New Local Politics of Water* (Cambridge, MA: MIT Press)

Olsaretti S 2018 Introduction: the Idea of Distributive Justice *The Oxford Handbook of Distributive Justice* ed S Olsaretti

US Gen. Account. Off. 1983 *Siting of Hazardous Waste Landfills and Their Correlation with Racial and Economic Status of Surrounding Communities* (Washington, DC: US Gov. Print. Off)

USEPA 2018 Interim OECA guidance for enhancing regional-state planning and communication on compliance assurance work in authorized States (available at: www.epa.gov/sites/production/files/2018-01/documents/guidance-enhancingregionalstatecommunicationoncompliance.pdf)