Chapter 2

Data

2.1 Raw Data

The Risk Screening Environmental Indicators (RSEI) Model is a geographically detailed data set produced by the Environmental Protection Agency (EPA). RSEI data is based upon the Toxic Release Inventory (TRI), which (for about 30 years) has been collecting data on all toxic releases in the US. The TRI program manages the regulations, policies and facilities that ultimately are mandated to be reported.

TRI data is self reported by facilities, with each observation being a release of a reporting chemical at a reporting facility. For each observation, data is collected on which chemical was released, how much of it was released, and the facility from which it was released.

The detailed location and chemical data that is collected through TRI (as well as detailed weather data from NOAA) is reformatted through a fate and transport model to create RSEI. The RSEI data shows where each release has traveled on an 800m grid across the USA. The ultimate data we have access to through the RSEI data is an observation for each release, for each square in the grid that the release hits. This gives us an idea of how the chemicals spread from the release locations, and enables us to create a map across the entire nation for where TRI chemicals accumulate in any year between 1988 and 2014.

The initial data from RSEI is in the form of one observation per release per block on the 800m grid that it reached. Given that the 800m grid across the United States has on the order of 10 billion squares, each release hits many squares, we have many releases, and the data covers 27 years, this is an enormous data set. Overall, the disaggregated microdata is approximately 4 terabytes of data.

The RSEI reformat of the TRI provides some additional information computed from the release information, and additional geographic information. X and Y are the geographic identifiers for the square on the grid across the US, with (0, 0) in the

center of the US. Release number tells us which release that row is associated with. Chemical number through media are all release specific data, but all data beyond that is release and grid number specific. Conc (concentration) is the raw concentration of the amount of the chemical that reached that grid cell. ToxConc is the toxicity weighted concentration of that release in that grid cell. The various 'Score' variables are meant to be used as hazard created by that release in that grid cell, as they are weighted by the population in that grid cell.

Below see an example of the disaggregated microdata:

X	Y	Release	Chem	Facility	Media	Conc	ToxConc	Score	SCancer	SNoCan	Pop
-185	51	2050156	317	3	1	4.55e-4	2.28e-3	0	0	0	0
-184	41	2050156	317	3	1	3.29e-4	1.65e-3	0	0	0	0
-184	42	2050156	317	3	1	3.33e-4	1.67e-3	0	0	0	0
-184	43	2050156	317	3	1	3.33e-4	1.66e-3	0	0	0	0
-184	44	2050156	317	3	1	3.35e-4	1.68e-3	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...

2.2 Important Caveats

Due to the nature of the data, there are a few interesting caveats to consider.

- The RSEI data is self reported, and has been thought to contain some severe underreporting.
- The data is entirely based off a black box fate and transport model. The model has uncertainty that we are not addressing.
- The data only captures releases for certain facility types within certain industries, for certain chemical types within those facilities. Not all chemicals are mandated reporting, and any analysis that is done based off the data can't be extrapolated to discuss toxicity more generally.
- Not only does the data not capture all chemicals, it also doesn't adress many common nuicances. Because of this, it is difficult to relate the RSEI scores to health or quality of life outcomes in an area. In discussion of environmental toxicity more generally, it is worth noting that certain toxicity types may be more or less likely to cause adcerse effects based on how they reach people in the area (eg. they may or may not reach local water systems). There are also more obvious environmental hazards, that are likely to have strong influence on public health and living conditions that TRI doesn't address, eg: brownfields, solid waste disposal, animal farming, hazardous waste, etc.
- RSEI gives the weight of the release and the chemical of the release, but chemicals have very different levels of toxicity. A small amount of mercury released is

much more problematic than a small amount of CO2. To that end, the EPA assigned each chemical a 'toxicity' weight, by which the amount of the chemical released is multiplied. This means that we can aggregate all the chemicals in an area, and compare the overall toxicity over space and time. The toxicity weights allow us to compare to other chemicals' toxicity weight, and overall toxicity of a given chemical release, but also means that the values only have meaning in comparison.

Despite the limitations that the data presents, it provides an incredibly detailed and complex view of toxicity in America that is worth delving in to.

2.3 Consistency in Reporting (detailed data description)

2.3.1 Census Comparability

For the time period that RSEI data is available for (1988 - 2014), Census geographies have experienced considerable overhaul. Many of our questions of interest involve demographic characteristics, and the changes we see in environmental toxicity over time for those demographics. As such, we need to be able to aggregate the toxicity to block group or tract level to be able to merge with Census data. To do so, we caluclate the toxicity values for each geographic unit for the closest temporal census geography (1990, 2000 or 2010). If we want to use Census areas as the unit of analysis, we also need to consider consistency of the units definition over time. To do so we would need to create crosswalks that help us transform past Census geographies to the current form, allocating the population appropriately to the new geographies.

2.3.2 Details of Chemical Consistency

Comparability across time, space, and chemicals is a consistent topic through this section. The EPA is incredibly detailed in their data collection, and creation of metrics to make the data meaningful on a broad scale. Unfortunately the EPA is subject to the changing scientific consensus of the times, and therefore hasn't been able to provide entirely consistent data.

Over time, as chemicals have been found to be toxic, they have been added to the list of TRI mandated reporting chemicals. There are also chemicals that through renewed understanding have been removed from the mandated reporting list. Because the list of chemicals reported changes over time, aggregating all the data would cause us to see artificial jumps in toxicity. These jumps wouldn't be reflective of an actual increase in toxicity, but rather that the toxicity was beginning to be measured. These jumps may change what areas appear as toxic i the data, as industries that don't have

to report at some point in the time frame will be entirely removed from the dataset. Several of those industries are very highly polluting, areas who focus strongly on those industries will show inaccurate low scores.

2.3.3 Details of Industry Reporting Consistency

TRI regulates who needs to self report using the North American Industry Classification System (NAICS), and before NAICS was available used its predecessor, the Standard Industrial Classification (SIC) system. Just as we see changes in the regulations for chemicals, we see changes in the regulations of various industries. NAICS codes that need to report are regulated independently of chemical codes, and NAICS codes that are not consistently reported across the time period of interest must be removed to maintain continuity. For a toy example, a textiles facility releasing mercury might have to report it, but the neighboring mining facility might not have to report their mercury emissions. If that changes over the time period, and suddenly mining needs to report, we will see an artificial huge jump in the mercury present in that area if we don't remove by industry.

2.4 Ultimate Data Form

The disaggregated microdata from RSEI is one observation per release per block on the 800m grid that it reached. The munging for these 4 terabytes of data filters each of the billions of observations to check that it is 1) from a chemical that is consistent across the relevant years 2) from an released that is linked to an industry that is consistent across years, and 3) allocates the observation to the appropriate geography.

This data cleaning is done through R, using the DBI and SQLite packages. Since there is so much data, it's not feasible to process it using typical R function, so after loading the data in to a database, it becomes queryable using SQL. This significantly speeds up the processes detailed below.

To accomplish the chemical consistency, we use a data table (provided by Rich Puchalsky) that contains a row for each chemical with the data of regulation, and the date of deregulation. Using this information, we can select the chemicals that are relevant to any set of years of interest. Chemicals are found by selecting the subset of chemicals where the year of initial regulation is before the interest period, and the year of deregulation is after the interest period, while also excluding delisted chemicals, and all observations are checked to be in this range.

Filtering out observations whose releases are not under a regulated NAICS category for the entire time is more complex. Using a similar table that contains the regulation and deregulation dates of NAICS codes we can find the consistent industry categories. However, the only reference to NAICS or SIC codes are in the facility table that the releases reference. This table provides 6 NAICS codes that are the most common

NAICS codes associated with the facility. However, NAICS codes are release specific, not facility specific, meaning that for each emission reported a NAICS code is reported. Removing by facility is not accurate, since facilities might have different types of NAICS emissions. The textiles facility we used as an example earlier might make both shoes and jackets, with different industry codes and different releases that have different reporting requirements. To get data on the NAICS codes by submission, data must be taken from the original TRI data, and linked to the disaggregated microdata by the document control number.

In the data cleansing process, we filter the data to remove industries and chemicals that were not consistent over the period of inquiry, and aggregate the data to the relevant Census geographies. The data, then forms a data set for each year, with an observation for each block group, and just one measure, an aggregated toxicity level.

block	concentration	area
010010201001	627.3050138	6.520168
010010201002	499.6297799	8.48669
010010202001	578.8311689	3.137173
010010202002	756.3733114	1.962949
010010203001	637.7356488	5.907125
...	...	...

The 'concentration' estimate must be interpreted in the context of the consistency adjustments. For example, towns with extremely high mining emissions may not show as exceptionally high pollution, as mining is one of the industries that had different regulations over the 1990-2010 time period, and therefore the reported mining emissions have been removed entirely. Essentially, the pollution estimates are comparable across time and location, but only in the context of continuous EPA regulation, and can not be interpreted independently of one another.

Chapter 3

Approaches and Methods

3.1 Geographic Level of Analysis

The questions we would like to pose - how the distributions of toxicity that individuals experience over time are predicted by their complex, multidimensional identities - is inherently intended to be a person level analysis. That intent may not be achievable given the available data.

This analysis depends on two data sources, the disaggregated RSEI toxic release data (as compiled to contain only releases that are consistently reported between 1990 and 2010) as well as relevant demographic information from the Census.

RSEI toxicity data can be obtained at extremely fine level (the 800 meter grid across the United States,) but the finest grain Census data is available at is the block level, which contains between 0 and a few hundred people. At such low geographic levels, few variables are available for demographics due to identifiability concerns. At low levels cross tabulations are not available due to concerns of identifiability. Using a low level of geography (like census blocks or block groups) is important for the environmental aspect of this analysis, since environmental hazards can be very localized, especially along neighborhood lines in urban areas.

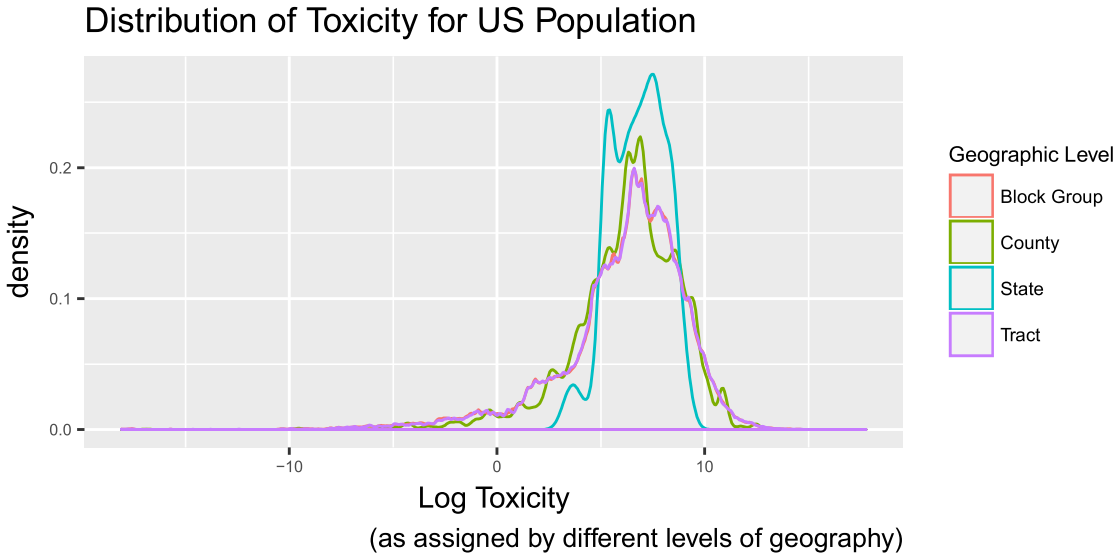
Unfortunately, the availability of cross tabulations is equally important to the goal of this work in examining inequality of environmental burden held by minority groups in America. The intersection of social identies, especially those steeped in systems of oppression, is extremely important for identifying unequal burdens. For example, low income populations across the board may be more likely to experience environmental hazards, but low income minority populations could be much more likely than low income white populations to experience extreme hazard. The intersections of demographic charictersitics, such as race and income or race and education are likely to be important in teasing out differences true inequality burden.

We combine the computed aggregated toxicity for each block group and the demo-

graphic data. Now for each census geography, we have toxicity information as well as demographic data.

block	concentration	area	total_pop	white	black
010010201001	627.3050	6.520168	530	447	83
010010201002	499.6298	8.486690	1282	1099	126
010010202001	578.8312	3.137173	1274	363	824
010010202002	756.3733	1.962949	944	458	477
010010203001	637.7356	5.907125	2538	2152	384
...	...	...	...	...	...

To move to a person level analysis, we can assign each of the people the toxicity for the geography they originated in. By assigning each person the toxicity of their block group, we can aggregate nationally to find the distribution of toxicity that each group experiences. This approach is restricted in ability to approach the problem in an intersectional manner, since we can only build a distribution for each of the crosstabs we have available. For higher levels of geography (where we might, for example, have race by income) we would be able to build national distributions for each income by race group. In the case of the table above, to build a distribution for the white population, we would assing 447 people a toxicity of ~627, 1099 people a toxicity of ~499 and so on until we have the full distribution of toxicities experienced by the white population. In choosing the level of aggregation at which to assign toxicity, in order to balance the needs of accuracy of toxicity and availability of cross tabulations, we create the overall toxicity distribution for Americans at each of the levels of geography. The process described above can be excecuted with the data shown above, or at a cruder level of geography, such as state. Using block group as the smallest form of geography, and state as the largest (including tract and county in between) we see how the distribution changes at each level of aggregation.



As expected, the state level assignment is a poor approximation of the lower level assignments. Given that we are assigning each individual the mean toxicity in their entire state, we are eliminating most of the variation from the data. Interestingly the tract data seems to build a distribution very similar to the block group level assignment. This may be because the block group level is aggregating a large enough group of our fine grain toxicity data that it lost the street block by street block variation that we had deemed so crucial, so aggregating several block groups gives us a conceptual equivalent 'neighborhood' level of aggregation.

3.2 Non-Parametric Analysis

To examine the change of environmental burden over time for groups we first use a non-parametric simulation technique to tease apart the forces in play as each group's distribution changes over time. For any given value, the percentile it holds in a minority distribution is likely to be different from the percentile it holds in the overall distribution. For example, in 2010 the 75th percentile of the black distribution is 4734.07, while that same value is the 80.74th percentile in the overall distribution.

We expect the mean of minority distributions to reduce over time for two reasons: first we assume the entire distribution will slowly be shifting right as we see improvements in environmentally friendly production technology and improved envirnmental regulation. Secondly, we hope that with Title VI protetections, and the work of civil rights advocates, minority communities will better protected against the economic power frequently held by polluters.

In order to track how minority distribution and the overall distribution have changed over the period of study, we use the positions that minority groups held in the overall distribution at the start of the period of study to simulate how each group's distribution would have progressed through time assuming a static position in society at large. This simulation proceeds as follows:

- Build an emipirical distribution of toxicity experienced for the entire population and for groups of interest in the starting year.
- Sample individuals from the empirical distributions for the entire population and the groups of interest.
- For each sampled value, find the percentile in the empirical distribution for the entire population in the starting year.
- Create an empirical distribution for the entire population in the ending year.
- For each sampled percentile, find the corresponding value in the full empirical distribution of the ending year and assign to the appropriate group of interest.

Using this method we can hold constant the place each individual (and more broadly each group) holds in the overall distribution, but follow the changes in the distribution

as a whole. The collection of values simulated now represents where each individual or group would have been had there been no positional improvement for the group as a whole.

If there had been improvement for a group, we would expect the simulated distributions to paint a much bleaker picture of the environmental burden borne than the true distribution of the ending year.

3.3 Temporal Level of Analysis

Given the layout of this non-parametric method, we can find the changes in positional distribution for any two given years. Though we are most interested in the complete change from the starting year of 1988 to 2014, as that is the data we have, the gradual changes and the speed of change from year to year is also of interest. While we have the full range of data from 1988 to 2014 for the toxicity data, we only have available Census data from the decennial census and the ACS. That means that we have snapshots of data from 1990, 2000, and continuous data for 2010 on.

3.4 Trends Over Time