Examining Injustice in Environmental Toxicity

A Thesis
Presented to
Department of Statistical Science
Duke University

Anne Driscoll

May 2018

Approved for the
Bachelor of Science in Statistical Science

_____

David Banks

_____

Committeemember O. Name

_____

Committeemember T. Name

_____

Mine Cetinkaya Rundel, DUS

Acknowledgements

I want to thank a few people.

Preface

This is an example of a thesis setup to use the reed thesis document class (for LaTeX) and the R bookdown package, in general.

Table of Contents

List of Tables

List of Figures

Abstract

The preface pretty much says it all.

Second paragraph of abstract starts here.

Dedication

You can have a dedication here if you wish.

Introduction

Environmental justice literature has focused largely on the motivations and policy approaches to help account for the disproportionate burden of environmental hazards on minority communities. The questions being asked are critical to building a knowledge base that allows us to create policy that helps correct the problem of hazard allocation, but aren't significantly addressing the canges in the hazard burden over time as it relates to policy already in place and natural sorting.

This thesis seeks to address the investigative question of if minority communities have experienced substantial reductions in their relative environmental toxicity, when those shifts are occuring, and a look in to the traits that are predictive of high environmental toxicity.

0.1 Keywords

Environmental Inequality, Racial Inequality, Non-parametric analysis

Chapter 1

Introduction to Environmental Justice

testing testing

123

Chapter 2

Data

2.1 Raw Data

The Risk Screening Environmental Indicators (RSEI) Model is a very detailed data set produced by the Environmental Protection Agency (EPA). It is based upon the Toxic Release Inventory (TRI), which (for about 30 years) has been. The TRI program manages the regulations, policies and facilities that are ultimately reported.

TRI data is self reported, with each observation being a release of a reporting facility for a reporting chemical. For each observation, data is collected on which chemical was released, how much of it was released, and the specific 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 across an 800m grid across the USA to create the RSEI data. The ultimate data we have access to through the RSEI data is an observation for each release, for each square in the grid 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 are spreading at any time between 1988 and 2014.

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 from certain industries, for certain facility types within those 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 broach many common nuicances, both enviromental and toxic. Because of this, it is difficult to relate the RSEI scores to health outcomes in an area. In discussion of hte toxicity, it is worth noting that certain toxicity types may not present themselves as easily (eg. 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 as chemicals have very different levels of toxicity. To that end, the EPA assigned each chemical a 'toxicity' weight, by which the amount of the chemical 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 locations, 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 during (1988 - 2014), Census geographies have experienced considerable overhaul. Many of our questions of interest involve demographics, 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. As such, we caluclate the toxicity values for each geographic unit for the closest temporal census geography (1990, 2000 or 2010).

2.3.2 Details of Chemical Consistency

Using the consistent_chemicals table provided by Rich Puchalsky, the chemicals that are relevant to the specific years of interest are calculated in the runDatabase.R file in the define_consistent_chemicals function. They 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.

2.3.3 Details of NAICS Code Consistency

NAICS codes are a form of industry code. In the RSEI data the only NAICS code reference is in the facility data table. This table provides a list of 6 NAICS codes that are most relevant to the facility. However, NAICS codes are release specific, not facility specific, meaning that for each emission reported a NAICS code is reported. Since NAICS codes are regulated independently of chemical codes, 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. 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. Removing by facility is also not accurate, since facilities might have different types of NAICS emissions. The textiles facility might make both shoes and jackets, with different industry codes and therefore 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 microdata by the submission and release tables. The Document Control Number (doc_ctrl_num) and Industry Code (industry_code) columns from the Submission NAICS Codes (v_tri_submission_naics) table provide the information necessary to link to our microdata.

2.4 Ultimate Data Form

The initial data from RSEI is in the form of one observation per release per block on the 800m grid that it hit. 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.

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, after an extensive and computationally intesive process, forms a data set with an observation for each year, for each block group/tract, with just a toxicity level.

The pollution 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.

Chapter 3

Approaches and Methods

3.1 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 environmental 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. Additionally, few cross tabulations are 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.

To combine the data, we assign the aggregated releases to their respective geographies.

For each census geography, we also have estimates of number of various population groups. From there, we move to a person level analysis by assigning 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.

~~insert the look at how the distributions differ when aggregated at each census level.

3.2 Need to get census data for income, education, and cross tabs on race and income.

3.3 Need to figure out how to do a time series analysis. do I want to do a

3.4 I want to do a non-parametric simulation of what would have happened to the distributions if they had stayed in the same place as they were in the original distribution.

How does the comparison there work? Do I just say wow look the real new distribution is actually significantly better, showing that there is real improvement for people of color. If we break it up by the states that implemented laws regarding environmental justice do we see a significant difference in the difference in differences?

AKAA Step 1) simulate the new distribution and see how the new and old dists. compare to the simulated new dist. Step 2) do we see that hte new is better than the simulated new? Step 3) do we see the same "positional advancement" in the states with and without laws regarding ej? Step 4) can we test that there is a difference in "positional advancement" (aka difference in simulated and real means) Step 5) can we test that there is a difference in "positional advancement" for the states that have good general environmental laws? Conclusions: there has been an improvement in the position of poc in the american toxicity distribution, and it is not/may be related to the specific ej laws.

????? what time period do I want to do that over? This is an iterational process,

~~right? I can't just do one block of time (eg 1990 to 2010) so how do I incorporate all~~ the jumps in time? ????? how can I do a longitudinal linear analysis? I would have a toxicity at each point, with the trend predicted by both the contents of the tract and the changes in the tract contents.

```
vars = c("white", "black", "hispanic", "owners", "renters")
p1990 = t1990
p1990[vars] = p1990[vars]/p1990$tot_pop
p1990$density = p1990$tot_pop/p1990$area
m = lm(lconcentration ~ 1 + owners + renters + black + hispanic + density, data =

p2010 = t2010
p2010[vars] = p2010[vars]/p2010$tot_pop
p2010$density = p2010$tot_pop/p2010$area
m = lm(lconcentration ~ 1 + owners + renters + black + hispanic + density, data =
```

Chapter 4

Results

testing testing

123

Conclusion

testing testing

123 hello

Appendix A

Code Appendix

This appendix includes all of the R chunks of code that were hidden throughout the document (using the `include = FALSE` chunk tag) to help with readibility and/or setup.

In the main Rmd file

```
# This chunk ensures that the thesisdowndss package is
# installed and loaded. This thesisdowndss package includes
# the template files for the thesis.
if(!require(devtools))
  install.packages("devtools", repos = "http://cran.rstudio.com")
if(!require(thesisdowndss))
  devtools::install_github("mine-cetinkaya-rundel/thesisdowndss")
library(thesisdowndss)
```

In Chapter 3:

References

Angel, E. (2000). *Interactive computer graphics : A top-down approach with opengl*. Boston, MA: Addison Wesley Longman.

Angel, E. (2001a). *Batch-file computer graphics : A bottom-up approach with quicktime*. Boston, MA: Wesley Addison Longman.

Angel, E. (2001b). *Test second book by angel*. Boston, MA: Wesley Addison Longman.