Choropleth Visualization of EPA Superfund Sites

Connie Lin and Matthew Windsor
Drexel University

Abstract—Current mapping techiques used by TOXMAP Environmental Health e-Maps regarding Superfund sites are little more than static points points on a thematic map. While this method is useful for navigational purposes, the ability to overlay corresponding data such as cancer rates or childhood asthma is limited, as this data is often presented in other formats. The Centers for Disease Control and Prevention's Division of Cancer Prevention and Controls is promoting the use of choropleth mapping as the defacto standard for planning and evaluation of cancer rate research and cancer control programs. One of the primary concerns related to EPA Superfund Sites is the health effects of toxic releases. We are presenting a three part choropleth visualization to correspond with the current datasets,spatial analysis and spatial statistic techniques endorsed by the CDC.

Index Terms—Information filters, industrial polution, data visualization,geospatial analysis.

1 Introduction

Maps are a useful and effective instrument for information visualization and evaluation. Visual information contained on maps offers a marked level of expediency over written data sets when determining spatial relationships or areas of interest. This expediency can be further enhanced by tailoring the map to a specific purpose, such as a topographical or political map. When comparing two sets of data in a visual context it is important to match the proxy information parameters to produce a true representation of the data. Scale, methodology, context and labelling are important considerations when comparing data sets.

The Environmental Protection Agency (EPA) designated the term Superfund as the environmental program established to address abandoned hazardous waste sites. This term is also used as the name of the fund established by the Comprehensive Environmental Response, Compensation and Liability Act of 1980. While the EPA policies and documentation are transparent regarding these projects, the information is fragmented. An independent project called Superfund 365 attempted to highlight the worst of the Superfund sites by visualizing the contaminants and responsible parties for each site. However, the entire set of sites is outdated and only represented with a sortable table, hiding trends in the dataset as a whole. [1] The EPA offers regional mapping but does not currently offer a comprehensive geospatial representation of the Superfund. The current standard for displaying Superfund geospatial information is the TOXMAP:® Environmental Health e-Maps by the National Library of Medicine (NLM) [2]. The visualization is static points on a thematic/political map of the continental United States (Figure 1).

Figure 1

The box representing the site is uniform and cannot be filtered. As the size is arbitrary, the information available to the website user is little more than a vague area. In the case of dense concentrations this vague area is often obscured by surrounding data points, making even a rough visual estimate impractical.

Lacking an existing practical evaluative visualization, we chose to focus on health evaluation measures, a primary concern regarding EPA Superfund Sites as related to the effects of the toxic releases. A literature review yielded many articles regarding choropleth map design as related to heath planning and evaluation. The Centers for Disease Control and Prevention's Division of Cancer Prevention and Controls is promoting the use of choropleth mapping as the de-facto standard for planning and evaluation of cancer rate research and cancer control programs [3]. Following the CDC recommendations regarding schemes, labels, projections and formatting we compiled Superfund data to closely match the current techniques used by the National Program of Cancer Registries (NPCR) and the National Cancer Institute (NCI).

2 Methodology

2.1 Data Gathering

This project gathered datasets from the CERCLIS Public Access Database [4], the Hazard Ranking System (HRS Toolbox) [5], TOXMAP:® Environmental Health e-Maps [2],Toxics Release Inventory (TRI) [6] County Cross Reference File (FIPS/ZIP4) from the Center for Disease Control [7]. The CERCLIS Public Access Database was used to gather the EPA ID, site name, city, state and county information. The EPA ID was determined to be a unique identifier and was used as a key for compiling additional data. The Superfund Chemical Data Matrix (SCDM) to determine the Hazard Ranking System score. The HRS uses mathematical equations for determining the relative threat posed by a hazardous waste site and reflect hazardous substance characteristics, such as toxicity and persistence in the environment, substance mobility and the potential for bioaccumulation including screening concentration benchmarks, environment or health-based substance concentration limits [8].

As many Superfund Sites were in the final stages of clean-up and had a zero HRS, we chose to select the top five hundred sites, as this number only included three zero level sites. TOXMAP:® Environmental Health e-Maps was used to ascertain street level locations. TOXMAP provides GIS visualizations of Superfund sites, but is unable to overlay any additional data, such as HRS or filtering capabilities. Using the EPA ID to align the datasets, the street level data was added to the final dataset. Google Maps Engine was then used to determine latitude and longitude. The address was concatenated into a single column with the gx_location label. Since several addresses contained internal commas, all strings were set in double quotes. The latitude and longitude was then extracted from the XML. The raw XML data was converted to plain text data and find/replace functionality of NotePad++ was used to eliminate extraneous formatting. An opening/closing root element was added to the plain text to convert the data to excel. This data was concatenated with the previous GIS data. The conversion dataset County Cross Reference File (FIPS/ZIP4) from the Center for Disease Control was used to determine FIPS (Federal Information Processing Standards) county 10-4 numerical reference.

2.2 Data Normalization

As mentioned in the data gathering section, the primary dataset was derived from multiple government agencies and concatenated using the EPA ID as the primary key. The primary dataset covered the EPA_ID, name, street level address, city, state, zip, county, FIPS, longitude, latitude, status, and HRS score. This data was then normalized into three separate tab separated value files. The Superfund Sites per County visualization found in Section 3.1 incorporates the frequency of sites in a given FIPS area as the primary tuple. The Highest HRS by County (Section 3.2) visualization combined both the FIPS and the HRS score. The highest value HRS in a given FIPS area was retained. The FIPS and the highest HRS score provided the tuple for this visualization. The third visualization, Combined HRS, added the HRS values for all sites occurring in a given FIPS. These values were then concatenated into a FIPS and summative HRS tuple.

2.3 Data visualization

We used the choropleth maps from the D3.js library developed by Mike Bostock to visualize the three sets of Superfund sites. Short for Data-Driven Documents, D3, was made for manipulating and visualizing data using Javascript, HTML, SVG, and CSS. [9] The choropleth maps presented in this paper are adapted from the choropleth map provided by Mike Bostock that uses quantize scale to display unemployment rates by county. [10] We used the JSON file referenced on the tutorial page to create the map of the United States broken down by state and county. Since the JSON file referenced each area by its FIPS code, we were able to map our Superfund sites to the relevant area.

For each map, the file and attribute names were updated to match our data. The color scheme was changed from blue to red, with the darker hues mapped to higher HRS scores or number of Superfund sites. For the number of sites, the domain in the quartize method was set to 0 to 10, as the maximum number of sites a single county contained was 9. For the visualization of the highest HRS score for a Superfund site by county, the domain was set to 0 to 100 to match the possible values of the HRS scoring system. [11] The map of combined HRS scores by county uses the threshold instead of the quartize method. Since the range of combined HRS values is arbitrary and influenced by the number of Superfund sites in the county, it made sense to use a method that allowed for arbitrary scaling. [12] The number of delineations was also decreased from 9 to 5 for this map. The legend included in each map was also adapted from a tutorial provided by Mike Bostock, specifically the one accompanying his threshold map of Kentucky population density by county. [13]

During development, our team utilized git and GitHub for revision control and sharing, respectively. Our public repository can be found at the following URL: https://github.com/GoodEveningMiss/info633-projectD. A working example is currently hosted using GitHub Pages at http://goodeveningmiss.github.io/info633-projectD/sites.html.

3 Visualization

Using the data we gathered for the 500 highest HRS scoring Superfund sites, we constructed three choropleth sites. Larger versions of the visualizations are provided in the Appendix.

3.1 Superfund Sites per County

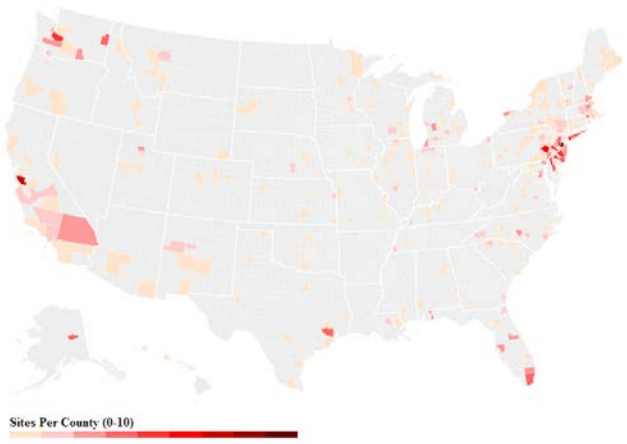


Figure 2

We mapped 331 different counties with at least one Superfund site. The county with the highest number of Superfund sites is Middlesex, New Jersey with 9 sites, followed by Santa Clara, California with 7.

Table 1: Breakdown of Sites per County

Number of Superfund Sites	Number of Counties
1	236
2	54
3	22
4	8
5	6
6	2
7	1
8	0
9	1

3.2 Highest HRS by County

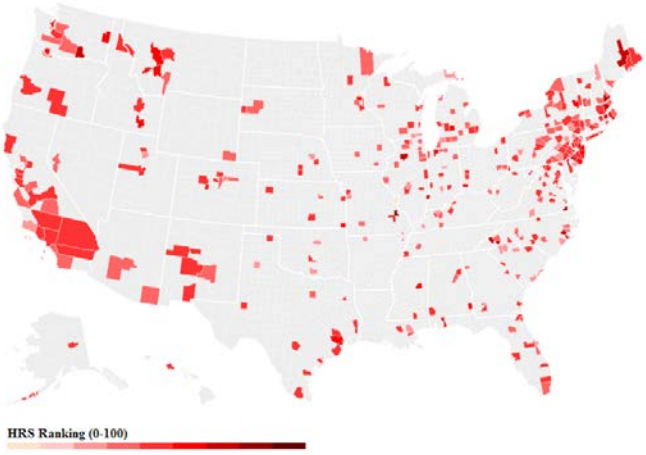


Figure 3

As stated in Methodology, the HRS system ranges is normalized to 100-point scale. Thus, the Scale for this choropleth map is from 0 to 100. Our dataset contained 3 sites with a HRS score 0, and the highest observed HRS is 84.91. Thus the colors representing the highest ranges are not represented in the Figure 3.

3.3 Combined HRS by County

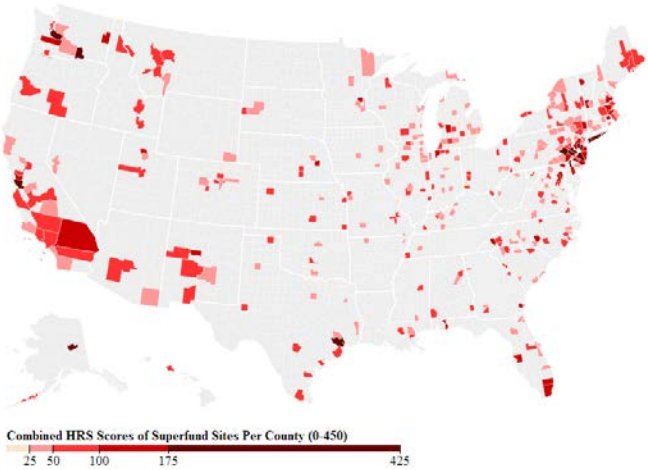


Figure 4

Not surprisingly, the county with the highest number combined HRS ranking is the county with the highest number of Superfund sites, Middlesex, New Jersey.

4 Discussion

4.1 Findings

4.1.1 Superfund Sites per County Visualization

The Superfund Sites per County visualization (Figure 2) represents the frequency of sites within a single FIPS (county level). This visualization is primarily used to compare the combined HRS visualization in regard to comparing total number of sites relative to clusters of high HRS sites. Although there is little change between the two visualizations, it shows clustering alone is representational of high toxicity scores.

Through the visualization, we can see larger trends in Superfund data. The counties with more Superfund sites tend to be centered on the Mid-Atlantic area and Southern California. While the geographic trends are interesting, it is difficult to determine from the map alone the contributing factors. Certain areas may have more Superfund activity due to early colonization and development, before environmental laws were enacted and enforced in earnest. Population density and presence of nearby waterways are also likely contributing factors.

4.1.2 Highest HRS by County Visualization

The highest visualization (figure 3.2) represents the highest HRS site within a single FIPS (county level. This visualization was used to show a “hot point” within a specific FIPS. This visualization could then be compared to the combined HRS visualization to show if a single site was primarily responsible for a high level of toxicity within a given FIPS. The findings actually showed the opposite. While the highest visualization share the same base level of sites, the combined scores did not indicate a single site was primarily responsible for a high combined FIPS.

4.1.3 Combined Visualization

The combined visualization (figure 3.3) represents the combined HRS scores for all sites contained within a single FIPS. This visualization is our most theoretical, as the HRS score is primarily used to represent single sites. The HRS is based on likelihood that a site has released or has the potential to release hazardous substances into the environment, characteristics of the waste (e.g. toxicity and waste quantity); and people or sensitive environments (targets) affected by the release) it makes sense that multiple sites within a close geographical area represent a greater threat than isolated areas. The results from this visualization were the most striking. Several counties are prominently displayed as “hot spots” with combined levels that

greatly exceed the majority of the national averages. This visualization may also show a “cross contamination” of sites (pollutant spread overlapping areas) that would not be visible in the traditional data point approach.

4.2 Weaknesses

Unfortunately, the exact coordinates of Superfund site boundaries is difficult to find. The choropleth map is not a one to one representation of the size of the associated superfund site or of the present toxic release. Since the data is bounded by the county the HRS scale becomes indicative of the entire county. The use of other visualization tools were considered, such as heat mapping using the latitude and longtitude of a site.

The map is not historical. Previous sites and the cumulative effects within aquifers, reservoirs, and air masses are not quantified or represented.

The combined map uses a hybrid approach regarding the HRS scale. The HRS is based on likelihood that a site has released or has the potential to release hazardous substances into the environment, toxicity and waste quantity; and people or sensitive environments (targets) affected by the release. It is logical to conclude that multiple sites within a close geographical area represent a greater threat to the human population than isolated areas. This is strictly a theoretical approach and is independent of the HRS methodology.

Puerto Rico was included in the datasets but was not represented in the visualizations. This functionality could be incorporated into future iterations.

4.1 Future Work and Applications

Due to the time constraints, only certain features and functionalities could be developed. As mentioned in the Introduction, development of these maps could have be useful when applied to the health sciences. It allows for easy overlay of additional datasets, and could be used to highlight areas in potential need of specific health-related services or possible areas of study. Aggregating Superfund data further to the state level and setting time constraints on Superfund sites shown may reveal areas where environmental policies and enforcement need strengthening.

Additional features such as and live filtering for a single dynamic map versus three static maps could improve the range and utility of these visualizations. Our dataset also includes much more information than is currently displayed, and the implementation of tooltips on mouseover would be an effective way to reveal more information to users while maintaining a low barrier to use. Our data is open and available to the public on our GitHub repository and individuals are welcome to contribute to further development.

5 Conclusion

Given the complex nature of evaluation of public health research, such as cancer rate research and cancer control programs, the environmental factors need to be analyzed with the same approach as the NPCR in determining incidence data in a GIS setting. Choropleth maps are a “common starting point” for mapping and analyzing incidence data [3]. The current mapping solutions provided by the EPA, CDC and National Library of Medicine do not include this functionality. Moreover, the datasets provided by the EPA require significant manipulation to conform to the NPCR standards. We hope that this research will lead to policy changes within the EPA in regard to data transparency and practical visualizations.

6 Acknowledgments

The authors wish to thank the U.S. EPA for making the Superfund site data available, Mike Bostock for the development, release, and documentation of the D3.js library and choropleth source code.

7 References

[1] "Index of Sites," [Online]. Available: http://turbulence.org/Works/superfund/about.html.

[2] NIH, "TOXMAP:® Environmental Health e-Maps," [Online]. Available: http://toxmap.nlm.nih.gov/toxmap/main/index.jsp. [Accessed 5 March 2014].

[3] T. Richards, Z. Berkowitz and C. Thomas, "Choropleth Map Design," *Preventing Chronic Disease, Public Health Research, Practice and Policy,* vol. 7, no. 1, p. 6, 2010.

[4] EPA, 3 March 2012. [Online]. Available: http://www.epa.gov/superfund/sites/cursites/. [Accessed 4 March 2014].

[5] EPA, "Introduction to the Hazard Ranking System (HRS)," 15 June 2013. [Online]. Available: http://www.epa.gov/superfund/programs/npl_hrs/hrsint.htm. [Accessed 3 March 2014].

[6] EPA, "Toxics Release Inventory (TRI) Program," [Online]. Available: http://www2.epa.gov/toxics-release-inventory-tri-program. [Accessed 5 March 2014].

[7] CDC, "County Cross Reference File (FIPS/ZIP4)," [Online]. Available: http://wonder.cdc.gov/wonder/sci_data/codes/fips/type_txt/cntyxref.asp. [Accessed 5 March 2014].

[8] EPA, "Superfund Chemical Data Matrix (SCDM)," 26 February 2014. [Online]. Available: http://www.epa.gov/superfund/sites/npl/hrsres/tools/scdm.htm. [Accessed 2 March 2014].

[9] Bostock, Mike. "Data-Driven Documents." *D3.js*. N.p., n.d. Web. 16 Mar. 2014. <http://d3js.org/>.

[10] Bostock, Mike. "Choropleth." Bl.ocks.org. N.p., n.d.

Web. 16 Mar. 2014. <http://bl.ocks.org/mbostock/4060606>.

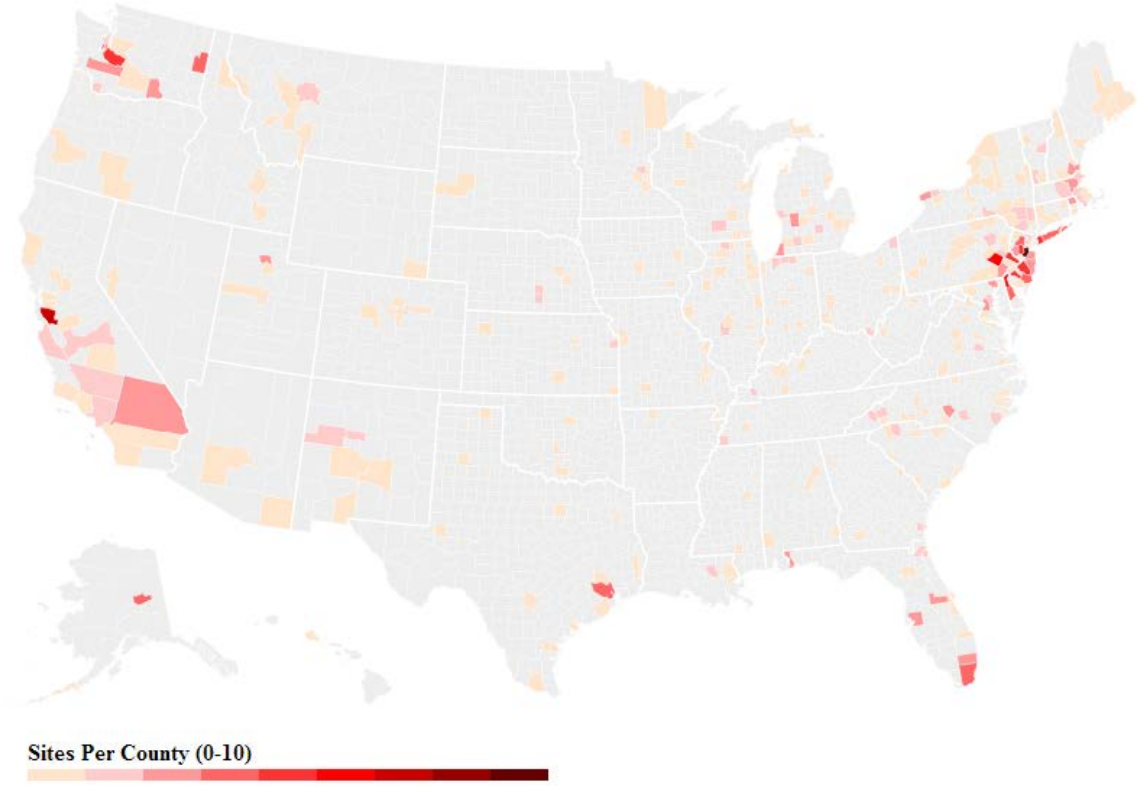
[11] "Chapter 1: Introduction." *Hazard Ranking System Guidance Manual*. N.p.: n.p., n.d. N. pag. *HRS Toolbox | National Priorities List*. U.S. EPA. Web. 16 Mar. 2014. <http://www.epa.gov/superfund/sites/npl/hrsres/hrsgm/ch1.pdf>.

[12] Cukier, Jerome. "D3: Scales, and Color." Jerome Cukier: Communicating with Data. Jerome Cukier, 11 Aug. 2011. Web. 16 Mar. 2014. <http://www.jeromecukier.net/blog/2011/08/11/d3-scales-and-color/>.

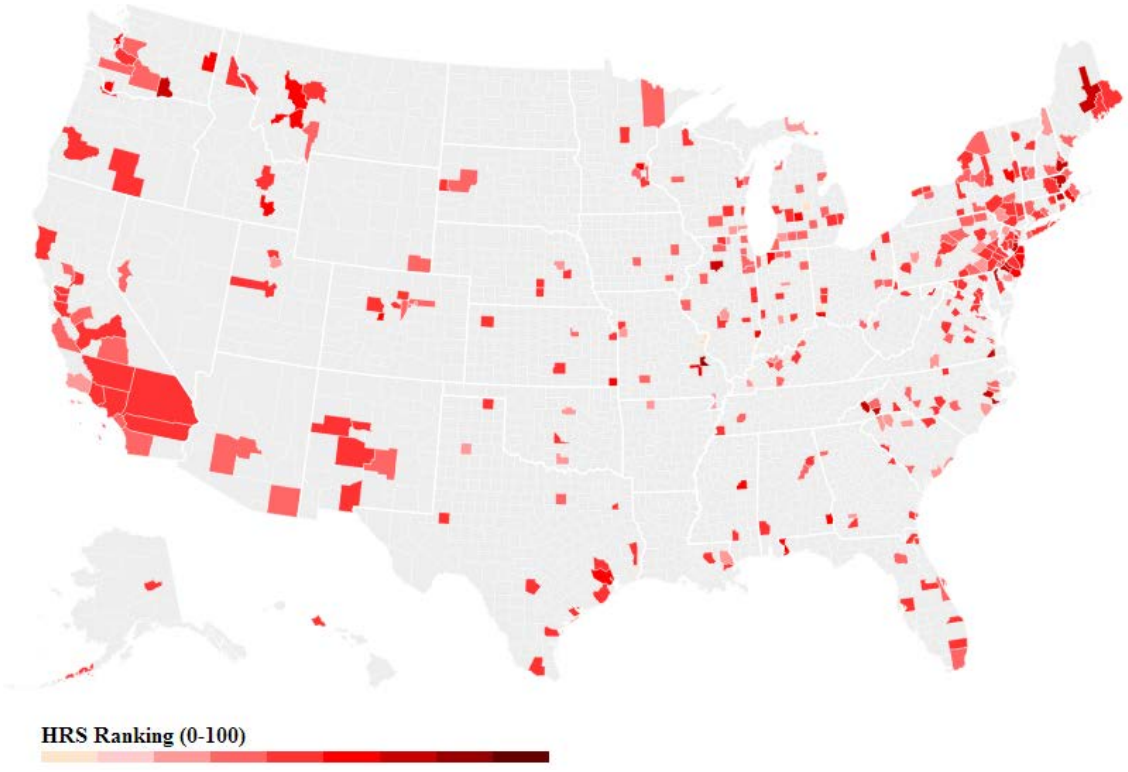
[13] Bostock, Mike. "Kentucky Population Density." *Bl.ocks.org*. Mike Bostock, n.d. Web. 16 Mar. 2014. <http://bl.ocks.org/mbostock/5144735>.

8 APPENDIX

8.1 Number of Superfund Sites by County



8.2 Highest HRS Score by County



8.3 Combined HRS Score by County