Proceedings of the 3rd World Congress on New Technologies (NewTech'17)
Rome, Italy – June 6 – 8, 2017
Paper No. ICEPR 107
ISSN: 2369-8128
DOI: 10.11159/icepr17.107

A Cluster Analysis of USA Superfund Sites and Associated Cancer Risk

Raid Amin, Arlene Nelson, Shannon McDougall
University of West Florida
11000 University Parkway, Pensacola, Florida 32514, USA
ramin@uwf.edu

Abstract **–** Superfund sites are geographic locations selected by the U.S. Environmental Protection Agency for long-term cleanup due to extreme toxic chemical spills. We study some characteristics of locations with Superfund sites. We have addressed three main research questions in this paper: (1) Are there geographical areas where the number (or density) of Superfund sites are significantly higher than in the rest of the USA? (2) Are there geographic areas where the overall cancer incidence rates are significantly higher than in the rest of the USA, and is there an association with the number (or density) of Superfund sites? (3) Are counties with Superfund sites more likely to have higher rates of minority populations than the rest of the USA? We studied the geographic distribution of overall cancer incidence rates (2007-2011), in addition to the geographic variation of Superfund sites for 2013. We used the disease surveillance software package SaTScan, which includes a scan statistic, to identify spatial clusters in cancer rates and in Superfund count and density. We also used the surveillance software FleXScan to support and complement the results obtained with SaTScan. Our results show that geographic areas with Superfund sites tend to have elevated cancer risk and elevated rates of minority populations.

Keywords**:** Cancers; chemical spills; scan test, environmental pollution

1. Introduction

Chemical spills can cause severe environmental and health problems. The Environmental Protection Agency (EPA) orchestrates the clean-up of some of the most out-of-control toxic sites in the country. After the EPA has been notified of a potential hazardous waste site, the site will undergo one or more assessments to determine whether it meets the criteria for clean-up. Sites determined to have the potential for serious hazards are then placed on the National Priorities List (NPL) and are eligible for Superfund Trust Fund-financed clean-up. The sites that are not cleaned up can pose serious health or environmental threats [1]. Some common contaminants found at Superfund sites included arsenic, lead, mercury, and polychlorinated biphenyls (PCB) [1]. These toxins, along with others, can impact surface water, groundwater, soil, air and even buildings [1]. One study found associations between excess cancer deaths from 1970-1979 and hazardous waste site locations in which the sole sources of ground drinking water were contaminated. The same study also found a cluster of elevated rates of gastrointestinal cancers in counties located in EPA Region 3 (Delaware, Maryland, Pennsylvania, Virginia, and West Virginia), an area with many Superfund sites [2]. There have been several Environmental Justice studies conducted to try to determine which segments of the population are most adversely affected by Superfund sites [3],[4],[5]. Most research suggests that non-white populations as well as Hispanic populations are more likely to live near Superfund sites [3],[4],[5]. Some studies have also found that areas with higher levels of poverty and lower levels of education may also be impacted [3],[4],[5]. However, the roles of race and ethnicity seem to be larger indicators than poverty and education [3],[4].

1.1. Statement of Problem

This paper aims to study the effect of Superfund sites on all 48 states in the lower (contiguous) United States of America. Specifically, we aim to answer the following questions:

1.Are there geographical areas where the numbers (or densities) of Superfund sites are significantly higher than in the rest of the USA?

2.Are there geographic areas where the overall cancer incidence rates are significantly higher than in the rest of the USA, and is there an association with the number (or density) of Superfund sites?

3.Are counties with Superfund sites more likely to have higher rates of minority populations than the rest of the USA?

It is important for county, state and even the national levels of government to be aware of the threats to people who are being most impacted by Superfund sites. Counties may need additional resources to be sure that they are adequately protecting their residents from the hazards associated with such pollution. Furthermore, people living near any hazardous waste sites have the implicit right to know whether their cancer risk is elevated due to some environmental factors. If a government knows that a particular site is associated with increased cancer risk, then steps may be taken to protect the health of the population. A map of Superfund sites in the United States can be found at: https://gispub.epa.gov/oeca/WOS/.

1.2. Data

The data for this study were obtained from (i) Environmental Protection Agency (EPA) [5], (ii) National Cancer Institute (NCI), (iii) Center for Disease Control (CDC)/Surveillance, Epidemiology, and End Results (SEER) Program [6], and (iv) United States Census Bureau. The Superfund site information was from the EPA's CERCLIS database. The data was current from the EPA as of November 2013, and included 1,699 sites for the 48 contiguous states. We included all hazardous waste sites that were proposed, included on, or declared by the EPA as "deleted" from the NPL. Data was the number of Superfund sites in each county of the 48 states being studied.

Cancer incidence data for the five-year period of 2007-2011 was reported as an average incidence count and age-adjusted rate per year. Kansas, Minnesota and Nevada did not report cancer incidence rates by county to the NCI. Therefore, the counties in these three states were not included in any analysis which included cancer incidence rates. In addition, the NCI did not report cancer incidence rates for twelve counties with small populations because these counties had less than sixteen occurrences of cancer over the five-year period, which is a set rule used by NCI for suppression of data. Counts for these counties were estimated using the overall state incidence rate and using a maximum of three occurrences per year. Cancer data was gathered for all types of cancer, all age groupings, and all races and ethnicities.

Data for high school drop-out rates and poverty rates were available for census years only. In order to obtain data for years 2007-2011, we performed a linear interpolation using the data from 2000 and 2010 to estimate values for 2009. County areas were obtained from the Census Bureau, as were county centroid coordinates and county adjacency information. The county size was used to calculate Superfund density, which we defined as the number of Superfund sites per 1,000 km^2 in a county.

1.3. Limitations

Our study was limited by gathering all data at the county level. Since not all Superfund sites affect large areas, some localized effects will not be found using data at this level. Census tract or zip code tabulation area data may be more effective at finding impacts on a smaller geographic level.

The cancer information from the NCI was reported as an average of five years of aggregated counts. While this should work well for review, it is possible that one year may have had an unusually high or low rate for one particular county. These spikes will not be apparent in this data. Furthermore, the NCI states in its data files that the individual state cancer registries may possess more local or current data [6]. Since gathering data from each state would have been a formidable task, we accept the limitation that our results are as accurate as the data the NCI has available.

2. Methods

2.1. Spatial Analysis

We used two spatial analysis tools, SaTScan™ v9.4.2 and FlexScan v3.1.2. SaTScan™ is a spatial and temporal analysis software developed by Martin Kulldorff, and used widely to study areas with disease epidemics [7]. SaTScan™ is capable of handling many different models, including Poisson, Bernoulli, and Normal. The spatial analysis in SaTScan™ allows the use of circular or elliptical windows with its scan statistics. We chose circular windows. For each location being analysed, the software scans a large number of circles ranging in size from a single geographic area to the upper limit defined by the model being run. FlexScan uses a nonparametric definition of clusters. It has a flexible spatial scan statistic developed by Kunihiko Takahashi and Toshiro Tango, which finds flexible-shaped clusters rather than just circular

clusters, by examining the geographical areas adjacent to each other [8]. Also, FlexScan includes a restricted likelihood ratio, developed by Tango, which takes each individual region's risk into account rather than looking at the cluster as a whole [8]. Unlike SaTScan™, FlexScan is limited to spatial analysis only and it does not contain an analysis for the normal model [9].

Superfund density (number of Superfund sites per 1,000 km^2) is a continuous variable, so we used the normal model in SaTScan™ to scan for clusters of high Superfund density. The normal probability model in SaTScan™ has a null hypothesis that all observations come from the same cluster, while the alternative hypothesis states that at least one cluster has a higher or lower mean than the area outside the cluster [10]. Clusters are identified by examining a large number of overlapping circles and determining a log likelihood ratio for each circle [10]. Superfund density was available for all 3,109 counties in the 48 contiguous states. The Superfund density was first normalized to a N(0,1) distribution in SAS v9.4. We then ran SaTScan™ using a case file containing the county code and Superfund density for each county. The location file was a file of county codes and the coordinates of county centroids. We used a purely spatial analysis, a circular scan statistic, and a maximum cluster size of 5% of the population. Since we are using the normal model, 5% of the population translates into 5% of the counties being analysed.

We examined cancer incidence data using a Poisson regression in SAS v9.4 to age-adjust the raw cancer counts. We then ran the Poisson model in FlexScan using both the restricted likelihood ratio and flexible scan statistic. We used alpha=0.2 for the restricted likelihood ratio and fifteen for the maximum cluster size, which are both defaults in the software. The analysis for cancer incidence clusters was run for 2,900 counties because cancer incidence rates were not available for three states. We adjusted the FlexScan county adjacency file to remove all references to counties in these three states so the cluster analysis would run correctly. The software treated the missing states in the same way it would treat any other separation between counties, such as a large lake. To examine the effects of the Superfund density on cancer incidence rates, we ran another Poisson regression with Superfund density added to the model. As before, we ran the Poisson model in FlexScan for the cluster analysis.

2.2. Other Statistical Tests

We used the Jonckheere-Terpstra (JT) test in SAS to study the relationship between the number of Superfund sites in a county and the cancer incidence rate of that county. The JT test is a nonparametric test that we used to test the following hypothesis: $H_0 : \mu_1 = \mu_2 = ... = \mu_k$ versus the alternative $H_1 : \mu_1 \leq \mu_2 \leq ... \leq \mu_k$ where at least one population mean (or median) is strictly less than at least one other population mean (median). This test may be done in this manner:

1.Arrange all data values in the order predicted.

2.Calculate P=how many values to the right are larger than the data value being considered.

3.Calculate Q= how many values to the right are smaller than the data value being considered.

4.The test statistic S=P-Q.

Superfund sites were considered by placing each county into one of five categories. Category one contained counties without any Superfund sites, categories two, three and four contained counties with one, two and three Superfund sites respectively, while category five contained counties with at least four Superfund sites. Twenty-two counties had ten or more Superfund sites. The cancer incidence rates were then compared across the five categories.

To determine the relative effects of several demographic variables on the make-up of a county containing a Superfund site, a stepwise logistic regression test was run in SAS v9.4. Of the 2,900 counties with cancer incidence rates, 710 contained at least one Superfund site. Counties with a Superfund site were assigned a value of 1, and those without a Superfund site were assigned a zero. The logistic regression was then run to determine the likelihood of a county containing a Superfund site based on the percentage of people in poverty, the percentage of African Americans, the percentage of Hispanics, the percentage of high school drop-outs, the cancer incidence rates, and the percentage of males in the county.

3. Results

3.1. Superfund Density

The Normal model in SaTScan found one significant cluster (with $p<0.05$) of high Superfund density in parts of Delaware, New Jersey, New York and Pennsylvania. There was also a secondary cluster in Virginia with $p = .069$. This

cluster is not significant, but since its relative risk is so high, it is an area that should be monitored. The map in Figure 1 shows both the significant cluster in yellow, and the secondary cluster in red. The details of the clusters are found in Table 1.

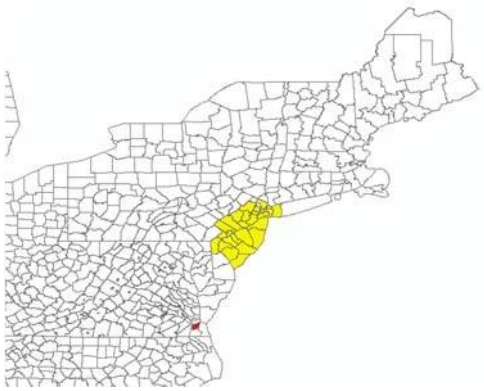

Fig. 1: Map of High Superfund Density Clusters in the United States.

Table 1: High Superfund Density Cluster Details.

States	Counties	Mean Inside Cluster	Mean Outside Cluster	PValue	Average Cancer Incidence Rate (per 100,000) of Counties	Average Poverty Percentage	Average African American Percentage
National Average					450.90 (Excl. KS, MN, NV)	16.21%	9.52%
DE,NJ, NY, PA	29 Counties	5.63	-.057	.001	499.39	11.37%	17.77%
VA	Norfolk, Portsmouth	11.78	-.0081	.069	479.95	17.86%	49.87%

3.2. Cancer Incidence and Superfund sites

To investigate the link between cancer incidence and Superfund sites, we first conducted a cluster analysis to identify the areas with higher-than-expected cancer incidence rates. We ran a Poisson regression with an age-adjustment for the raw cancer incidence counts. This adjustment removed the effects of age from each county's cancer incidence count, and we then used these results in a geographical cluster analysis to find areas where cancer incidence rates are truly higher than expected.

The results from our Poisson model with the restricted likelihood ratio run in FlexScan can be seen in Figure 2. The map shows all geographical clusters of increased cancer incidence. It must be noted here that Union County, Florida is included in a cancer incidence cluster and has a large prison population which could cause the county to show up as an outlier in disease studies [11]. The inmate population of the hospital is not counted in the Union County population at risk, but the cases diagnosed at the hospital are counted in the Union County cancer incidence [11]. The most likely cancer incidence cluster is the cluster with the highest likelihood ratio. It contains counties in Delaware, New Jersey, and Pennsylvania and has a relative risk of 1.171. This means the population within the cluster has a 17% higher risk of cancer than the rest of the U.S. Many of the counties in the most likely cluster are also found in the statistically significant cluster of increased Superfund density. Five of the clusters with the highest cancer incidence rates contained no Superfund sites at all, while the two clusters with the highest likelihood ratios contained significant numbers of Superfund sites.

Next, we adjusted the cancer incidence counts by both age and Superfund density and ran a second cluster analysis in FlexScan. The map shown in Figure 3 shows the effects of this additional adjustment. The counties in dark red were found in both the age-adjusted and age-Superfund density-adjusted models. The blue counties, which are detailed in Table 2, are those that were in an age-adjusted cluster, but not an age-Superfund density-adjusted cluster. These are the counties that have cancer incidence rates which can be attributed to Superfund density (listed in Table 2 as Counties Removed after Adjusting for Superfund Density). Finally, the light red counties are those that are only on the age-Superfund density adjusted results. Table 2 details the clusters that contain at least one county with high cancer incidence rates attributed to Superfund density.

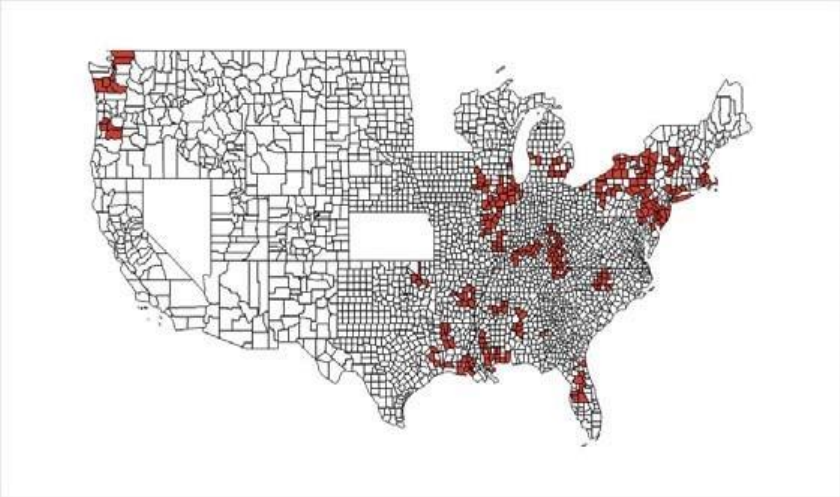

Fig. 2: Map of Age-Adjusted High Cancer Incidence Clusters.

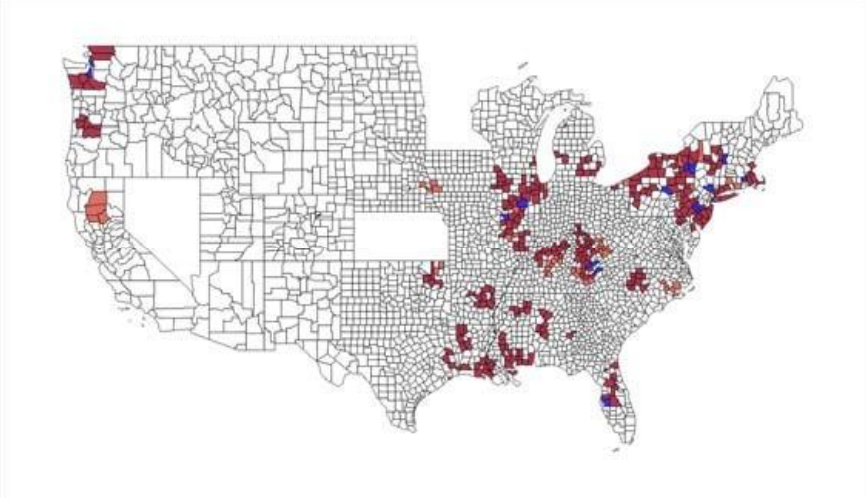

Fig. 3: Map of Age-adjusted and Superfund Density-adjusted Clusters.

Table 2: FlexScan Clusters Affected by Adjusting Cancer Incidence by Superfund Density (p<.01).

General Cluster Location	Number of Counties before Adjusting for Superfund Density	Relative Risk of Age-adjusted Cancer Incidence Cluster	Counties Removed after Adjusting for Superfund Density	Average SF Density of Removed Counties per 1,000 km^2
NY	7	1.102	Orange	1.8416
RI	7	1.081	Providence	7.0900
NY	12	1.123	Chenango, Cortland	0.5998
PA	9	1.067	Bucks, Montgomery	9.6160
IL	9	1.091	Morgan	0.0000
WA	8	1.094	Island, Kitsap	3.4753
FL	6	1.068	Hillsborough	5.1856
IL	8	1.146	McLean	0.0000
VT	9	1.111	Bennington	2.2790
PA	7	1.118	Clinton	0.4305
NY	4	1.034	Bronx, Kings, New York	2.6559
KY/TN	10	1.085	Bell, Harlan, Hamblen, Hawkins	0.2062

The Jonckheere-Terpstra (JT) test was used to test for the presence of a monotonic trend of cancer incidence rate increase as the number of Superfund sites increases. The test results are significant with a high test statistic (Z=8.5341) and p<.0001. The results show a definite trend of increasing cancer incidence rates, detailed in Table 3, as the number of Superfund sites in a county increases. The trend can also be seen in the graph shown in Figure 4. Each of the three quartiles (q25%, q50%, q75%) is monotonically increasing with increased number of Superfund sites per county.

Table 3: JT Test Results for Cancer Incidence and Number of Superfund Sites.

Number of Superfund Sites	Number of Counties	Average Cancer Incidence Rate per 100,000
0	2190	446.85
1	415	459.20
2	112	462.08
3	78	466.74
4+	105	478.90

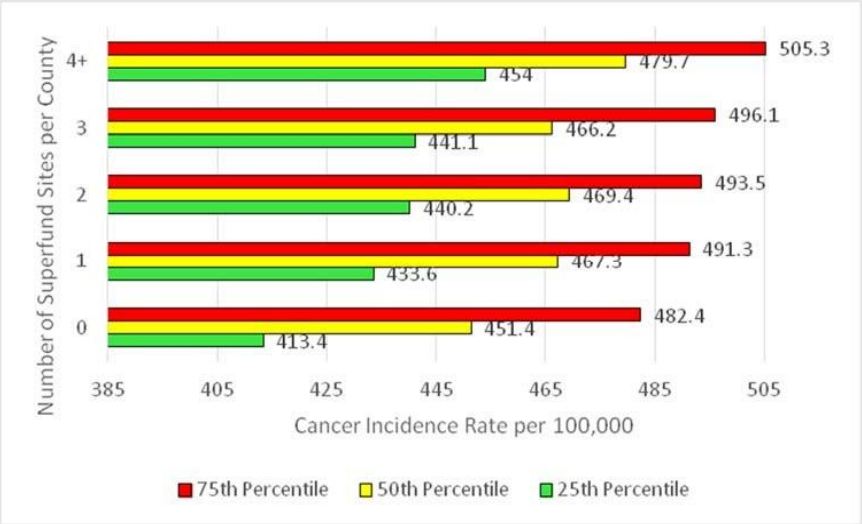

Fig. 4: Cancer Incidence Rates and Superfund Sites.

3.3. County Characteristics

The stepwise logistic regression included all variables in the model and had a concordant rate of 72.8% which is considered high. The odds ratios are shown in Table 4. With these results we can make several observations about counties with at least one Superfund site. They are likely to have more minorities, as measured by both race and ethnicity. The populations have higher levels of education, as measured by the high school drop-out rates. Interestingly, they are likely to have lower percentages of males. Cancer incidence and poverty rates did not significantly contribute anything to the model because their odds ratios were near a value of one.

Table 4: Odds Ratio Estimates from Step-wise Logistic Regression.

Effect	Point Estimate	95% Wald Confidence Limits	
Poverty	0.984	0.963	1.005
African American %	11.251	5.347	23.673
Hispanic %	113.236	49.308	260.048
High School Drop-out %	0.890	0.871	0.909
Cancer Incidence Rate	1.008	1.006	1.010
Male %	<.001	<.001	0.005

4. Conclusion

Superfund Density is the number of Superfund sites per 1,000 km^2. It measures the potential severity of a chemical spill relative to the corresponding county area. We identified clusters having both high cancer rates and high Superfund site densities. Our results show that geographic areas with Superfund sites tend to have elevated cancer risk, and also have elevated rates of minority populations. There may exist several important confounders. In particular, minorities may have less access to health care and thus higher rates of cancer. From an environmental justice standpoint, minorities may be more likely to live near Superfund sites; e.g., if the presence of a site depresses real estate values, then poorer people, often minorities, are likely to rent and own in that area. Regarding gender, aside from counties with prison populations, there is little variation, and this deserves deeper examination on a county by county basis.

References

[1] EPA, "Cleaning up the Nation's Hazardous Wastes Sites," Retrieved April 18, 2015 from United States Environmental Protection Agency's Learn the Issues, Science and Technology, [Online]. Available: http://www.epa.gov/superfund

[2] J. Griffith, R. C. Duncan, W. B. Riggan, and A. C. Pellom, *Archives of Environmental Health: An International Journal*, vol. 44, no. 2, EPA, "Report and Product Descriptions," Superfund Information Systems, 1989.

[3] J. T. Boer, M. Pastor Jr., J. L. Sadd, L. D. Snyder, "Is there Environmental Racism? The Demographics of Hazardous Waste in Los Angeles County," *Social Science Quarterly*, vol. 78, no. 4 pp. 793-810, 1997.

[4] K. Burwell-Naney, H. Zhang, A. Samantapudi, C. Jiang, L. R. Dalemarre, W. Lashanta, E. Williams, S. Wilson, "Spatial Disparity in the Distribution of Superfund Sites in South Carolina: an Ecological Study," *Environmental Health*, vol. 12, no. 96, 2013.

[5] P. Stretesky, and Michael J. Hogan, "An Analysis of Superfund Sites in Florida," *Social Problems*, vol. 45, no. 2 pp. 268-287, 1998.

[6] Surveillance, Epidemiology, and End Results (SEER) Program, SEER*Stat Database: State Cancer Registry and CDC's National Program of Cancer Registries, January 2014 Sub (1973-2011) <Katrina/Rita Population Adjustment> - Linked To County Attributes - Total U.S., 1969-2012 Counties, National Cancer Institute, DCCPS, Surveillance Research Program, Surveillance Systems Branch, released April 2014, based on the January 2014 submission. Accessed: December 2014, [Online], Available: www.seer.cancer.gov

[7] M. Kulldorff, "SaTScan User Guide," 2015, [Online]. Available: http://www.satscan.org/

[8] T. Tango, and K. Takahaski, "A Flexible Scan Statistic with a Restricted Likelihood Ratio for Detecting Disease Clusters," *Statistics in Medicine*, vol. 31., pp. 4207-4218, 2012.

[9] K. Takahashi, T. Yokoyama, T. Tango, "FlexScan User Guide for Version 3.0," [Online]. Available: http://www.niph.go.jp/soshiki/gijutsu/index_e.html

[10] M. Kulldorff, L. Huang, and K. Konty, "A Scan Statistic for Continuous Data Based on the Normal Probability Model," *International Journal of Health Geographics*, vol. 8, p. 58, 2009.

[11] C. Ren, S. Lim, T. Hylton, et al, Florida Annual Cancer Report: 2008 Incidence and Mortality. Tallahassee: Florida Department of Health, 2012.