GEOGRAPHIC ANALYSIS OF PUBLIC REQUESTS FOR CANCER CLUSTER INVESTIGATIONS AND USING IT AS AN INDICATOR OF HEALTH DISPARITY IN THE U.S

by

LINLI ZHU

(Under the Direction of Lan Mu)

ABSTRACT

Clusters of non-communicable diseases have become an important world-wide public health issue, especially cancer clusters. In the U.S, concerned individuals can contact state or local health agencies to report a suspected cancer cluster (SCC) for various reasons including perceived hazards. SCC investigation request, together with media exposure, can be a powerful alternative way to make voice heard. However, only less than one quarter of the requests get further investigation. This study emphasizes on the utilization of Geographical Information Science (GIScience) techniques and statistical methods to identify and understand disparities in those public concerned areas. The results of Montana reveal that cancer incidence rates in SCC areas tend to be lower than the rest; less disadvantaged population reside in SCC areas; and SCC areas associated with more environmental contamination.

INDEX WORDS: Suspected cancer cluster (SCC); Geographical Information Science (GIScience); Logistic regression; Health disparity; Environmental disparity

GEOGRAPHIC ANALYSIS OF PUBLIC REQUESTS FOR CANCER CLUSTER INVESTIGATIONS AND USING IT AS AN INDICATOR OF HEALTH DISPARITY IN THE U.S

by

LINLI ZHU

BS, Hunan Agriculture University, China, 2012

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment

of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2015

GEOGRAPHIC ANALYSIS OF PUBLIC REQUESTS FOR CANCER CLUSTER INVESTIGATIONS AND USING IT AS AN INDICATOR OF HEALTH DISPARITY IN THE U.S

by

LINLI ZHU

Major Professor: Lan Mu
Committee: Jerry Shannon
Xiaobai Yao

Electronic Version Approved:

Julie Coffield
Interim Dean of the Graduate School
The University of Georgia
May 2015

ACKNOWLEDGEMENTS

This thesis would not have been possible without the help of many people in so many ways. Firstly, I would like to express my gratitude to my advisor Dr. Lan Mu for the useful and detailed comments and suggestions on both research approaches and thesis writing. I would like to thank my committee members, Dr. Jerry Shannon and Dr. Xiaobai Yao, whose ideas help my research become more complete and robust. Also thank you Laura Williamson and Monawar Hosain for providing me suspected cancer cluster data. Last but not the least, I would like thank Lauren Anderson and UGA Writing Center for helping me on thesis writing.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Clusters of non-communicable diseases have become an important world-wide public health issue, especially cancer clusters. It is alarming to have a certain type of disease diagnosed frequently within a short time period and within a geographic area. Naturally, people would like to find causes and solutions to fight the disease cluster. In the United States (U.S.), concerned individuals can contact state or local health agencies to report a suspected cancer cluster (SCC) (National Cancer Institue 2014). From 1990 to 2013, the Centers for Disease Control and Prevention (CDC) issued three documents for guiding the investigations of disease clusters. Michael Goodman, et al in 2014 summarized those three documents: The first document was issued in 1990 and provided general guidelines for assessing clusters of health events. As used in these guidelines, “cluster” was defined as “unusual aggregation, real or perceived, of health events that are grouped together in time and space and that are reported to a health agency“ (CDC 1990). A four-stage approach for managing a reported cluster has been developed including initial contact and response, assessment, major feasibility study, and etiologic investigation (CDC 1990). From the perspective of public health, guidelines in 1990 emphasized the importance of the perception of a cluster in a community, which may be as important as an actual cluster. However, it did not specifically focus on cancer cluster investigations. In 2007, an addendum to the 1990 guidelines specifically highlighting investigations of cancer clusters was published (Goodman et al. 2014; Kingsley et al. 2007). In 2013, the CDC issued revised guidelines for investigating SCC and responding to community concerns, which continually discussed the four-step process for evaluating potential cancer clusters. The guidelines mainly concentrated in the scope of cancer clusters in a community, neither occupational cancer clusters nor medical treatment caused cancer clusters (Abrams et al. 2013). As we can see, there is a tendency that geographic cancer clusters are increasingly emphasized by the CDC.

According two surveys in 1989 and 1996, we can see an increasing number of SCC investigation cases (Thun and Sinks 2004). Each year, state and local health agencies receive over 1,000 inquiries regarding suspected cancer clusters in certain geographic areas, among which many citizens expressed perceived hazards, like an environmental contaminant, around their communities(Trumbo 2000). And For example, in May 2009, the Florida Department of Health received an email requesting assistance in investigating a possible increase in childhood brain cancers in an area west of Palm Beach called the Acreage, which stated that “there were an alarming number of brain cancers, particularly among children in her area”. For this inquiry, the first level of investigation was followed (Florida Department of Health 2009). However, few cluster concerns received further investigation due to the lack of personnel, resources and inadequate interagency communication (Juzych et al. 2007). According to Figure 1, up to three quarters of the inquiries (in blue) ended with initial response over the phone or in a letter from state to the informant or reporter. 5-30% typically involved an in-door review of mortality and incidence data, or examination of available environmental quality data (Greenberg and Wartenberg 1991). And only 1-3% of caller requests ended in further epidemiological investigations in the community, such as sampling of the environment. Identifying a causal relationship between a cancer cluster and an environmental toxin that is exposured to humans is particularly difficult to study mainly because of the complex etiology and the lack of residential history of patients.

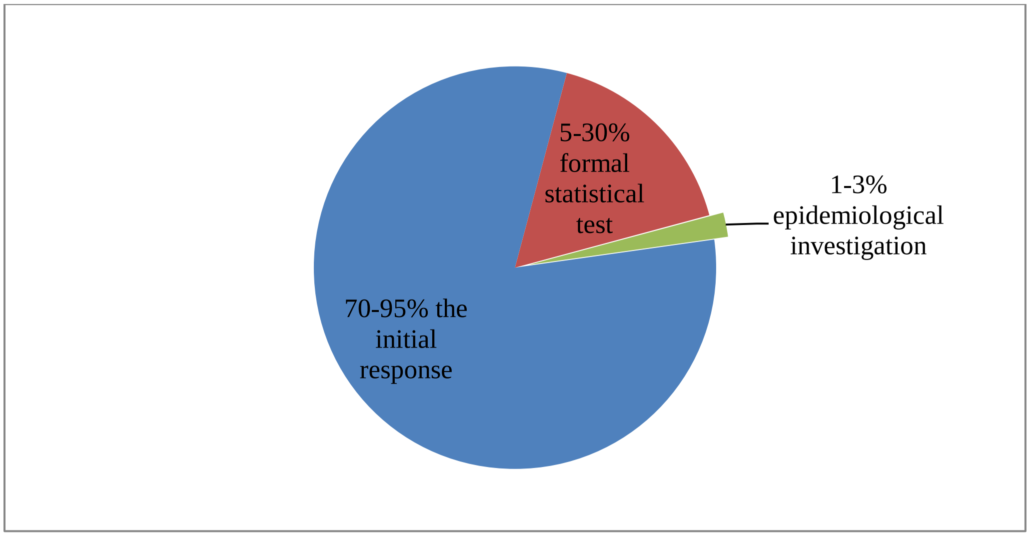


Figure 1 Investigation level of SCC reports

A SCC does not equal to an actual cancer cluster, but rather the perception of a cancer cluster. The perception level varies from individual to individual. Therefore, as demonstrated in Figure 2, a SCC area could be either confirmed as an actual cancer cluster area statistically through extensive epidemiology study, or not. Not surprisingly, actual cancer clusters may be located in non-SCC areas. Residents in these areas might be lack of awareness of cancer cluster or merely are not familiar with the procedure of reporting a SCC. Moreover, despite the conclusion by experts from health department, concerned individuals often persist in believing that the cancer cluster was not random (Siegrist 2001). Siegrist discussed that the persistence can be explained from two aspects, on one hand, people tend to identify patterns (and causes) instead of randomness, on the other hand, people are lack of social trust in public health experts. In many studies, social trust has been found to influence risk perception of hazardous waste, and have a significant influence on the perception of the cancer cluster (Siegrist 2001). Therefore, CDC tends to only focus on the process of identifying a statistical significant cancer cluster. However, under these guidelines and procedures, social awareness issue, risk perception issue and social trust issue might be hidden.

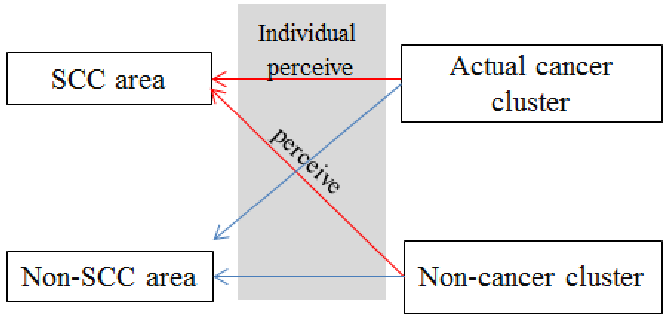


Figure 2 SCC conceptual model

Besides governmental health agencies, those with concern about cancer clusters or environmental contaminants can contact some non-profit organizations. For instance, consumer advocate Erin Brockovich’s Community Health Book is a platform for people to report contamination issues in their neighborhood or their observation of what appears to be excess illness. From this website, similar expressions like “there are excessive, excessive number of cancer cases in my hometown” and “I know way more people here with cancer or recovering from cancer” can easily be found (Community Health Book 2014). In California, there are total 9 reports that mentioned more and more people being diagnosed with cancer in their locations.

The press also published many cases of public concerns about SCC. In 2013, reported by Banger Daily News, a local newspaper in Maine, a citizen said that he knew of 15 people with some form of cancer living on the section of Coldrook Road and there were only 20 households within that span. The nearby landfill and several transportation companies have been mentioned by residents (Gagnon 2013). On January 7, 2001, St. Louis Post-Dispatch, a local media in Missouri published an article describing some O’Fallon residents believe their neighborhoods are witnessing many infant deaths and cancer cases including childhood leukemia, and their suspected cause is the federal cleanup of an old explosives and uranium complex at the Weldon

Spring. However, none of the cases have uncovered higher-than-expected rate of disease in the community (Shelton 2001). In an article from The Daily News (Jacksonville, North Carolina), Senior Scientist Gina Solomon of National Research Defense Council emphasized the importance of SCC investigations: “although it’s really difficult to conclusively prove that what caused any specific disease cluster, what I want to say today to you is that we can gather invaluable clues and hints from these events, and those together can help us solve the mystery of chronic disease" (Hodge 2011). Admittedly, investigations do not often establish clusters, but they can still help clean up the environment that can result in fewer cancer cases (Shelton 2001). However, recently, there is a group of scholars criticized that stories published on newspaper, sometimes, uncritically present cancer and local industry data, it transformed questionable statistics into an alarming public issue (Perez 2015).

In terms of health data collecting processes, the majority of data employed by public health related studies are always distributed by authorities in a top-down approach, which means that physicians or hospitals are required to submit healthcare data to health agencies and those agencies gather and process all the data and serve as a gateway to distribute the data . And I will refer this type of data as authoritative health data. For instance, there are two federal programs that support central cancer registries which receive and compile cancer cases from clinical facilities: CDC’s National Program of Cancer Registries (NPCR) and National Cancer Institute (NCI) Surveillance, Epidemiology and End Results Program (SEER) (Abrams et al. 2013). SEER covers approximately 28 percent of the U.S. population including many minority population registries (SEER n.d.). NPCR includes 45 states and represents 95% of the U.S. population (CDC 2013). Possessing these data, only some states agencies regularly analyze cancer registry data to identify communities with more cancer cases than expected (Thun and

Sinks 2004). Unlike this top-down approach, the records of public reports of SCC investigations are generated from a bottom-up approach that community members report concerns about SCC and potential environmental hazardous source, which is aptly in accordance with the concept of "citizens as sensors", proposed by geographer Michael Goodchild in 2007 (Goodchild 2007). Meanwhile, many scholars in the city planning field have proposed that we should bring local knowledge that community residents possess into environmental and health decision-making (Corburn 2007). Therefore, this study is aimed at listening to the voice from people, and taking some actions to investigate.

Unfortunately, analysis has been rarely conducted linking SCC geographic areas reported by the public to health disparity. On one hand, the data of those concerned geographic areas is not easily accessible and has not been systematically collected and organized in the U.S or even at the state level. On the other hand, health agencies are more likely to conduct traditional epidemiology studies emphasizing the causal connections, case by case, between specific environmental exposures and cancer site (Wakefield 2010). To my knowledge, no one has analyzed environment, health and disparity issues to assess those SCC investigation requests as a holistic research object. Looking at the big picture of public SCC areas such as a simple distribution map, is also a blank area.

Geographic Information System (GIS) techniques are used in this study, which is suitable to health and environment related studies because it allows for the integration of multiple data source, cartographic representation of data, and the application of various spatial analytical techniques for proximity analysis (Chakraborty 2011). Based on the outcomes from GIS analysis, statistic test and logistic regression are followed. In this paper, the modeling aims to discover the relationship between absence-presence of SCC investigations requests and cancer data collected from government, socioeconomic, and environment health factors.

Raising public awareness is a very important component in scientific research and policy making. Regarding SCC, the public serves not only as the data provider but also the research subject, and investigation results should be conveyed with them rather than kept within the academic circle. To my knowledge, we do not have a comprehensive information platform for sharing SCC reports and the investigation results with public. Therefore, tools such as an online web map application displaying a dynamic map of visualized public concerned geographic areas, which is accessible by everyone, is worth being created.

Building upon the knowledge gap identified above, my research question is: **what is the role of public request for cancer cluster investigations in understanding the geographic pattern of cancer and how to use it an indicator of health disparity in the U.S.?** . My research objectives are

1) Compiling a GIS dataset of public reported SCC in the U.S.,
2) Modeling and investigating SCC areas from the perspectives of health, environment and disparity based on geographical analysis.
3) Developing a web application as a platform to share SCC information with the public and provide broad impact of the thesis research.

CHAPTER 2

DATA COLLECTION AND COMPILATION

Up to this point, I have collected data from three State Health Departments, National Cancer Institute (NCI), Census Bureau, U.S. Geological Survey, and Environmental Protection Agency (EPA), and the following data introduction part is organized as three categories (Table 1): cancer data, demographic and socioeconomic data and environmental data.

Table 1 Data groups, source and period

Data Group	Data source	Year(s)
Cancer data	State Health Department	2006 - 2013
	State Cancer Profile	2007 - 2011
Demographic and socioeconomic data	Census Bureau	2010
Environmental data	EPA Toxic Releases Inventory	1987 - 2012
	EPA Impaired Water	2012
	USGS National Uranium Resources Evaluation (NURE) program	1975 - 1980

2.1 Cancer Data

At this point, public SCC reports to three states in the U.S have been collected and compiled. Firstly, SCC Investigations records published by Delaware Health and Social Service have been collected for this study. The reported geographic areas are at different scales. More than half of the concerned geographic areas are reported at the ZIP code level, and a couple of them were reported as Census Tract level. Data at county level are discarded in this study, because of the fact that there are only three counties in Delaware. Therefore, a list of ZIP codes and Census Tracts has been extracted as public concerned and SCC geographic areas from 2005 to 2009 (Table 2), and the corresponding geographic areas have been plotted on the map (Figure 3).

Table 2 Public concerned SCC geographic areas in Delaware

Year of Request	Geographic Area	Type of Cancer
2005	Zip code 19701	breast
2006	Zip code 19939	All types
2006	Zip code 19945	All types
2006	Zip code 19947	All types
2006	Zip code 19966	All types
2006	Zip code 19967	-
2006	Zip code 19970	All types
2006	Zip code 19975	All types
2007	Census tract 144.02	All types
2007	Zip code 19803	brain/CNS
2007	Zip code 19804	-
2007	Zip code 19810	breast
2007	Zip code 19970	All types
2008	Zip code 19808	breast
2008	Zip code 19810	All types/thyroid cancer
2008	Zip code 19904	All types
2008	Zip code 19958	Lung
2009	Census tract 104.00	Brain cancer
2009	Census tract 103	All types/ cancers in women
2009	Census tract 112.05	All types
2009	Census tract 502.00	All types
2009	Zip code 19939	All types/breast
2009	Zip code 19945	All types/breast
2009	Zip code 19947	All types/breast
2009	Zip code 19966	All types/breast
2009	Zip code 19970	All types/breast
2009	Zip code 19975	All types/breast

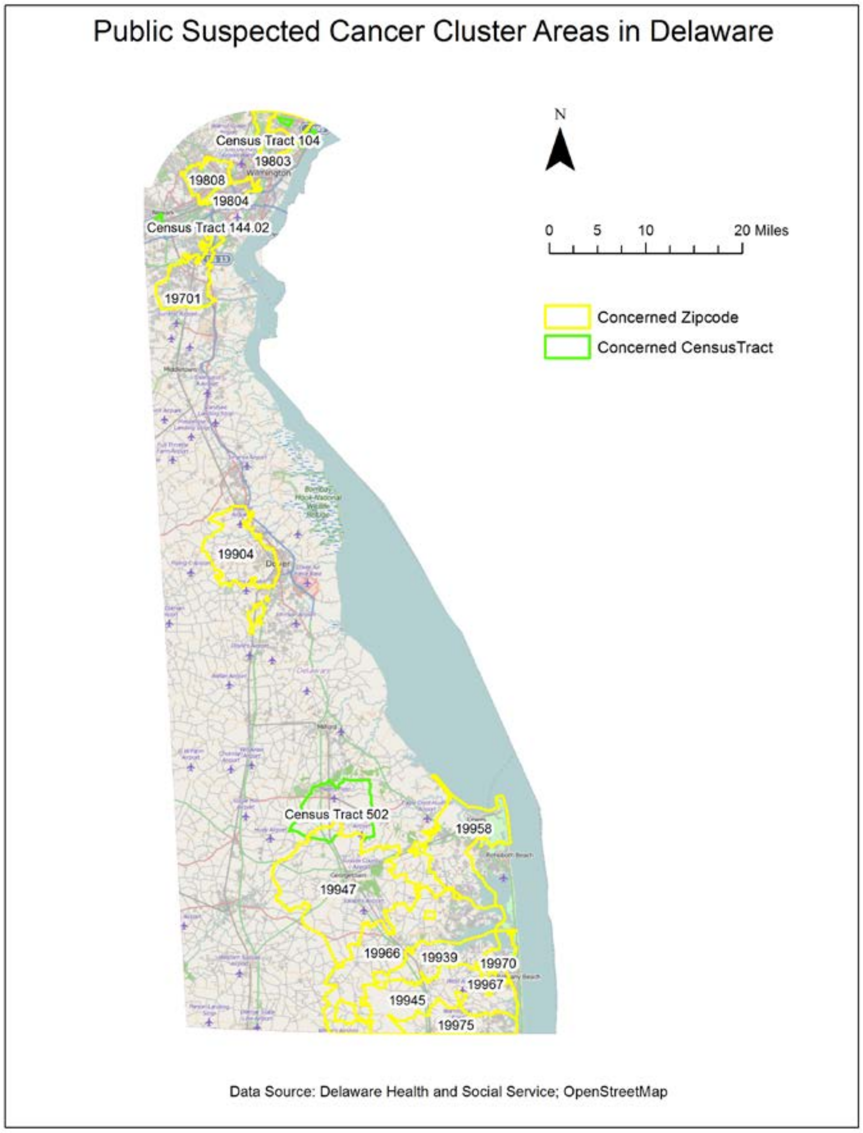


Figure 3 Map of public suspected cancer cluster areas in Delaware

To my knowledge, Delaware is the only one state whose SCC data is available and published online. In order to collect more SCC data, over twenty data requesting emails have been sent out to state health departments. New Hampshire Division of Public Health Services responded to the request by providing the SCC investigation requests in 2013 (Table 3). There were four town-level SCC reports in total. And Montana Department of Public Health and Human Services listed 22 cases from 2009 to 2013 whose geographic scales vary from city-level, county-level to state level (Table 4). For SCC data of New Hampshire and Montana, the distribution maps are also plotted (Figure 4, 5). From Figure 4 and 5, the majority of SCC areas are located along major roads and highly connected by the road system indicating that they are developed urban areas. And using TIGER/Line Shapefiles Urban Areas as reference, in Montana, more than half of the reported geographic areas are urban areas. This distribution is reasonable, since the majority of population reside in urban areas, so the probability that reports are from urban area increased. Geocoding and displaying all the public concerned geographic areas provide better spatial display, understanding and getting ready for GIS analysis.

According to U.S. Cancer Statistics in CDC, in terms of age-adjusted incidence rate combined all cancer sites, Montana, New Hampshire and Delaware ranked 23,40 and 47 out of 50 states respectively in 2011 (CDC).

Table 3 Public concerned SCC geographic areas in New Hampshire

Year of request	Geographic Area	Description
2013	Town of Rye, Rockingham County	A. 6 Female Breast CA B. 2 with Malignant Brain CA C. 1 with Meningioma
2013	Town of Bow	very rare childhood bone cancer
2013	Town of Gilford, Belknap County	Brain tumor
2013	Town of Merrimack, Hillsborough County	10 cancer cases in 60 houses

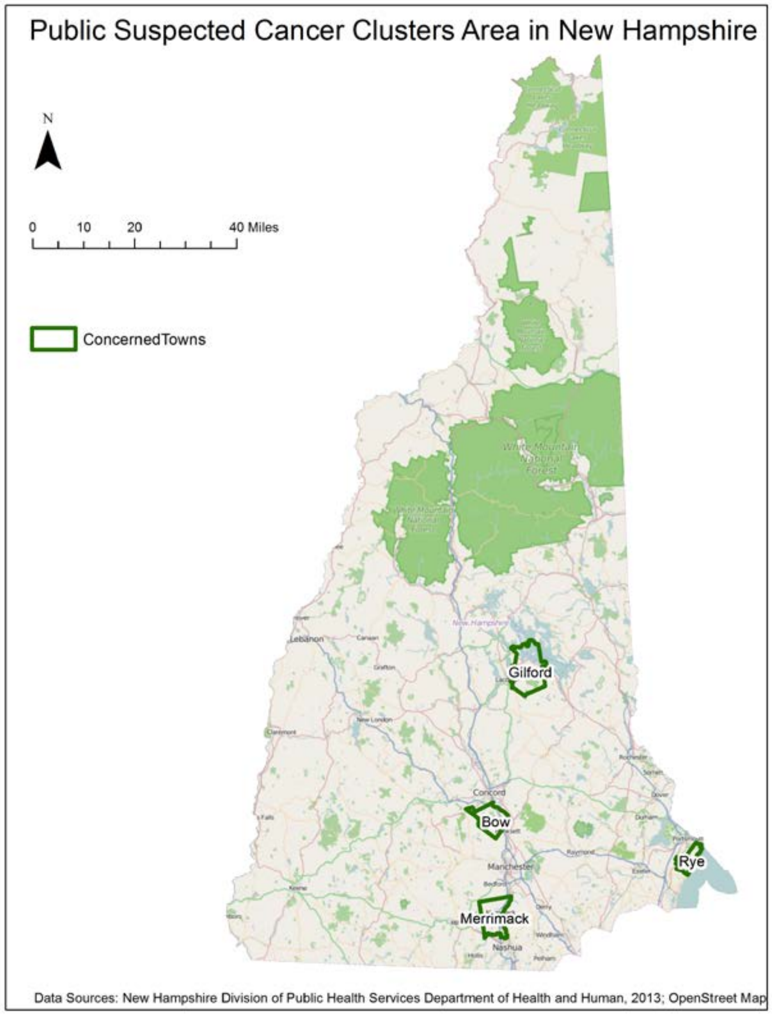


Figure 4 Map of public suspected cancer cluster areas in New Hampshire

Table 4 Public concerned SCC geographic areas in Montana

Year of request	Geographic Area	Geography Type
2009	Bozeman	city
2009	Butte	city
2009	Colstrip	city
2009	Fort Belknap Reservation	American Indian Reservation
2009	Glacier National Park	A particular building located within the park
2009	Kalispell	city
2009	Montana	state
2009	Montana City	city
2010	Butte/ Silver Bow County	city & county
2010	Colstrip	city
2010	Wolf Creek Canyon	A "neighborhood" (in a canyon) located in Lewis & Clark County
2011	Beaverhead County	county

2011	Florence	city
2011	Livingston	city
2011	Missoula	city
2012	Broadwater County	county
2012	Fort Peck Reservation	American Indian Reservation
2012	Kalispell	a particular business located in Kalispell
2012	Kalispell	city
2012	Noxon	city
2012	Silver Bow County	county
2013	Hot Springs	city
2013	Livingston	city

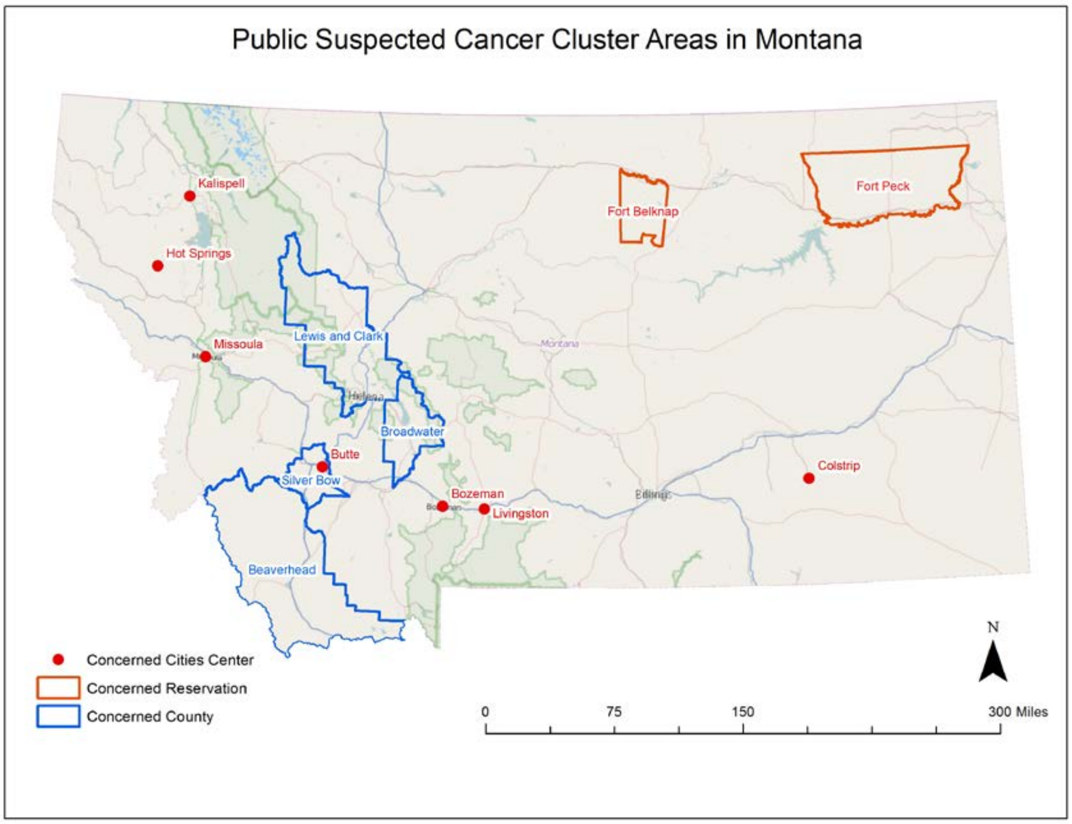


Figure 5 Map of public suspected cancer cluster areas in Montana

Authoritative cancer data includes age-adjusted cancer incidence and age-adjusted death rate at county level which are available on State Cancer Profiles (Abrams et al. 2013). The county level data is the finest scale that is accessible by public. This data provides a basic understanding of the distribution of cancer incidence rate and mortality rate (average from 2007 to 2011) in study areas, and thus comparison between data sources could be conducted (State Cancer Profiles). Additionally, high incidence and mortality rate could trigger the panics of cancer cluster from the public. My hypothesis on this is that high incidence and mortality rate have positive relationship with the emergence of public SCC reports, or it can predict SCC reports.

2.2 Socioeconomic and demographic data

Demographic data have been collected from the Census Bureau, which includes 2010 Census TIGER/Line shapefiles and the 2010 Census Summary File for the U.S. This data is used for calculating the ratio of different races, age-groups, median household income and percentage of population with bachelor degree among people 25 ages over in study areas. Because, in environmental justice field and health disparity filed, a lot of the research point that people with less privilege might disproportionately expose to hazards and have relatively worse health outcome, which could, consequently, lead to a concern of SCC. Incorporating Census Tract level data is important because, firstly, the size of Census Tract is generally match with many of the concerned geographic areas, secondly, neighborhood effects on health disparity analysis are emphasized heavily in studies, and Census Tracts are usually used as surrogates for neighborhoods due to their relatively homogeneous population characteristics, economic status and living conditions (Darden et al. 2010).

2.3 Environmental Data

Three categories of environmental data are collected from EPA: Toxic Release Inventory (TRI) data, water quality data, and air quality data. Because Trumbo's study summarized that 21% of the SCC investigation requests expressed a direct concern over some industrial situation;

water causes about 8%; others such as air quality and agriculture pollution for about 7%; also, concerns about radiation account for 3% of the SCC investigation requests (Trumbo 2000).

Firstly, the TRI program collects the amount of toxic chemicals that may pose a threat to human health and the environment released from U.S. industrial facilities, and it reports how much of each type of chemical is released to environment or has been safely treated (EPA 2015). TRI data is downloaded by building a search query in TRI.NET software specifying the geographic area and time period from 1987 to 2012; Occupational Safety and Health Administration (OSHA) defined Carcinogens chemicals are used as Filtering variables, and from which accumulated onsite releases to carcinogens chemicals can be obtained. The TRI sites with only on-site releases have been filtered based on a common assumption, which states that industries with only on-sites releases could potentially exert negative effects on people that live around them. Secondly, EPA's impaired water data is also incorporated and impaired water features reflect river segments, lakes, and estuaries. And this data can be used as a cross validation dataset for the following sediment sample dataset. Thirdly, for air quality, EPA's air quality nonattainment status as county level is incorporated. According to the International Agency for Research on Cancer (IARC), outdoor air pollution is classified as a cancer-causing agent (carcinogen). Therefore, air quality might be a determinate variable of public inquiries.

Another important dataset is obtained from USGS National Uranium Resources Evaluation (NURE) program, which collected sediment samples throughout the U.S and conducted geochemical analyses. Montana statewide database consists of 33979 sediment sample sites. In general, geochemical mapping can be considered as an important tool in environmental studies, and the objectives includes determination of geochemical background values and identification of natural or manmade chemical contamination (Zumlot 2009). For example, the Wolfson Geochmical Atlas of England and Wales is a geochemical mapping study developed in the 1960s. It consisted of over 50,000 stream sediment samples and helped to understand certain environmental issues such as diseases like molibdicosis in cattle, and others (Zumlot 2009; Howarth 1983). In the U.S. the NURE program sampled over 250,000 sites in the continental US and analyzed them for up to 40 constituents (Zumlot). There are also successful studies based on NURE data. For example, Ried (1993) developed a geochemical atlas of North Carolina using NURE data. Cocker (1999) conducted geochemical mapping in Georgia using NURE data as a tool for geological and environmental studies, and mineral exploration.

CHAPTER 3

METHODS

Because the objective of this study is to model and investigate SCC areas from the perspectives of health, environment and disparity, GIS techniques is used to integrate different types of data and perform spatial analyses. Table 5 summaries statistical methods employed in this research. Moran's I is used to test the spatial autocorrelation of cancer data at county level to test if counties with high cancer incidence or death rate are clustered, and then we can compare it with the distribution of SCC data. Based on some of the results of geospatial analyses, logistic regression model is performed aimed to discover the relationship between absence-presence of SCC investigations requests and cancer data collected from government, social economic, and environment health factors.

Table 5 Statistical methods summary

Statistical methods	Data	Scale	Purpose(s) of the method
Moran's I	Cancer Incidence rate Cancer death rate	County County	Test spatial autocorrelation, identify if high cancer rate clustered
T-test	All data	Census Tract	Investigate disparity between SCC areas and the rest
Correlation	All data	Census Tract	Explore association between collected data
Logistic regression	All data	Census Tract	Model absence-presence of SCC investigations

3.1 Data visualization and preliminary geographical analysis

Health disparity analysis can start with exploring and analyzing the distribution of cancer incidence rate and cancer death rate in the study areas by generating thematic maps and visualizing them. Quantitative analysis can then be followed to stochastically or deterministically investigating cancer clusters, of course, after tackling the inconsistent scale problem. Spatial areal interpolation method would be a solution.

The relationship between social economic position (SEP) and cancer is also explored as a reference to SCC. Racial group data at Census Tract level have been analyzed. 2010 Census data has been used to calculate white and black population ratios. Additionally, because 60 percent of all cancer occurs in persons over 65 years of age, and cancer incidence and death rates are also the highest in elderly age groups (Yancik 1997), ratios of population above 65 years old are calculated in order to examine inequality or disparity among elderly age groups. A series of distribution maps of Census Tract level SES variables have been generated using either Standard Deviation or Quantile classification methods to keep maps comparable, which allow us to identify the relatively vulnerable regions, such as places with a cancer death rate belong to the highest quantile and highest black population quantile.

In order to evaluate the potential hazardous influence of toxic chemicals on study areas, TRI data is used, to which Euclidian Distance and buffer analysis are applied. Output of Euclidian Distance surface, a distance raster indicating the distances from each cell to the nearest source locations (ArcGIS Resources 2012), can demonstrate the degree of potential impact from TRIs on every location. The accumulated toxic release from 1987 to 2012 are calculated for every TRI and also a total release amount in every Census Tract are obtained by Spatial Join. Furthermore, buffer analysis is worthwhile to analyze the geographical areas of public SCC in a more specific and quantitative way. The uncertainty about how to select the buffer distance on hazardous sites has been discussed in literature, and buffer analysis ranging from 100 to 1000 yard buffers have been experimented (Sheppard et al. 1999). In this study, both 5-mile and 1-

mile buffers are tested. Firstly, to deal with the edge effect, 5-mile buffering was created for each concerned geographic area in Montana and the number of TRI sites within this area or buffering zone has been calculated. Secondly, 1-mile buffer around TRI sites are computed. In the same manner, a 1-mile buffer on impaired water in Montana has also been calculated to represent proximity areas that could potentially be influenced by impaired water.

3.2 Geochemical data interpolation

Mapping the spatial distribution of NURE sediment samples requires spatial interpolation methods to generate continuous surface of chemical concentrations. Interpolation techniques such as such as Kriging and inverse distance weighting (IDW) have been extensively used in sediment and soil investigations and pollution mapping (Xie 2011). Interpolation techniques all have a smoothing effect, which will definitely lead to bias in geochemical mapping. All interpolation methods have been developed based on Tobler's first law of geography that everything is related to everything else, but near things are more related than distant things (Tobler 1970). IDW is based on the premise that the predictions are linear combination of available data. In IDW method, it is assumed substantially that the rate of correlations and similarities between neighbors is proportional to the distance between them and can be defined as a distance reverse function of every point from neighboring points (Yasrebi 2009). Ordinary kriging is one of the most basic of kriging methods. Both methods estimate values at unsampled locations based on the measurements at surrounding locations with certain assigned weights for each measurements (Yasrebi 2009). And there are many studies of the performance of the spatial interpolation methods, but the results are not clear-cut (Xie 2011; Shi 2009).

This study investigates the regional distribution of uranium, lead and zinc elements in stream sediment samples from Montana. I did not create a map for every chemical, because, on one hand, some chemicals' sample sites did not cover the entire study area, so the number of points is big not enough to perform spatial interpolation. On the other hand, the selected elements are harmful ones. For uranium, intakes of exceeding amount and lead to increased cancer risk or lead to internal irradiation and/or chemical toxicity (EPA). And the high lead values being mostly of anthropogenic origin (Lima 2003), which might indicate that hazards industries with emission of lead are nearby.

Besides looking at the elements separately, developing a synthesis map combining the effects of three types of chemicals would be more intuitive and can help us evaluate the exposure better. Consequently, Reclassify tool in ArcMap is used to assign every pixel a new score ranging from 1 to 10, the larger the value is, the higher the predicted chemical concentration is. And then, Raster Caculator can be conducted for combining all the interpolated surfaces. The equation is:

$$\mathrm{C} = \frac{\mathrm{Pb+U+Zn}}{3} \qquad (1)$$

Where C is the combined assessment of aforementioned three elements in sediment. On the right, Pb, U, and Zn are reclassified chemical values ranging from 1 to 10 respectively, and f or every element, they are equally weighted.

3.3 Modeling the presence of SCC reports

Logistic regression, also called a logit model, is used to model dichotomous outcome variables. In the logit model the log odds of the outcome is modeled as a linear combination of the predictor variables (IDRE).

The principle of logistic regression rests on the analysis of a problem, in which a result measured with dichotomous variables such as 0 and 1 or true and false, is determined from one or more independent factors (Menard 1995; Ayalew 2005). In the case of this SCC study, the goal of logistic regression would be to find the best fitting model to describe the relationship between the presence or absence of public SCC reports in a certain geographic area and a set of independent parameters such as cancer incidence rate, income, racial, and various environmental factors, etc. Generally, logistic regression involves fitting the dependent variable using an equation of the form (Ayalew 2005):

$$Y = \mathrm{Logit}(p) = \ln\left(\frac{p}{1-p}\right) = C_0 + C_1X_1 + C_2X_2 + \cdots + C_nX_n \qquad (2)$$

Where p is the probability that the dependent variable is 1 (Y: presence or absence of public SCC reports), $\frac{p}{1-p}$ is the odd ratio, C_0 is the intercept, and $C_1, C_2 \ldots C_n$ coefficient, which measures the contribution of independent factors $(X_1, X_2 \ldots X_n)$ to the variation in Y. The structure of the equation is similar with linear regression. In order to appropriately interpret the meaning of Eq. (2), one has to use the coefficients as a power to the natural log(e). The result represents the odds ratio or the probability that an event will occur divided by the probability that it fails to do so. If a coefficient is positive, its transformed log value will be greater than one, meaning that the event is more likely to occur. If a coefficient is negative, the latter will be less than one and the odds of the event occurring decreases. A coefficient of 0 has a transformed log value of 1, and it does not change the odds one way or the other. For a positive coefficient, the probability plotted against the values of an independent variable follows an S-shaped curve. A mirror image will be obtained for a negative coefficient (Menard 1995)

However, as mentioned, the reported geographic areas are at varies scales. To model the presence or absence of public SCC area, Census Tracts are grouped into two groups by the location of the central points to represent the SCC concerned and non-concerned areas. The criteria are that for those whose central point located in the concerned areas, they will be classified into concerned Census Tract and assigned value “1”. The rest would be the non-concerned Census Tract and assigned value of “0”.

Logistic model is only conducted for State of Montana, since the area of New Hampshire and Delaware are small, we do not have enough number Census Tracts as input to perform LR analysis. The glm() function in R software is used to build regression models, which fits generalized linear models, a class of models that includes logistic regression (James 2013). And vif() function in “car” package is used to test the multicollinearity.

CHAPTER 4

RESULTS AND DISCUSSION

This chapter introduces the results of statistical analysis and GIS based analysis of the aforementioned data. And the results are organized as follows: it starts with the results of data visualization in graph or maps formats, the discussion of geospatial analysis, and the results of visual tests and patterns are discussed; and then, using Montana as an example, the results of statistical test, correlation analysis and regression analysis are discussed more thoroughly.

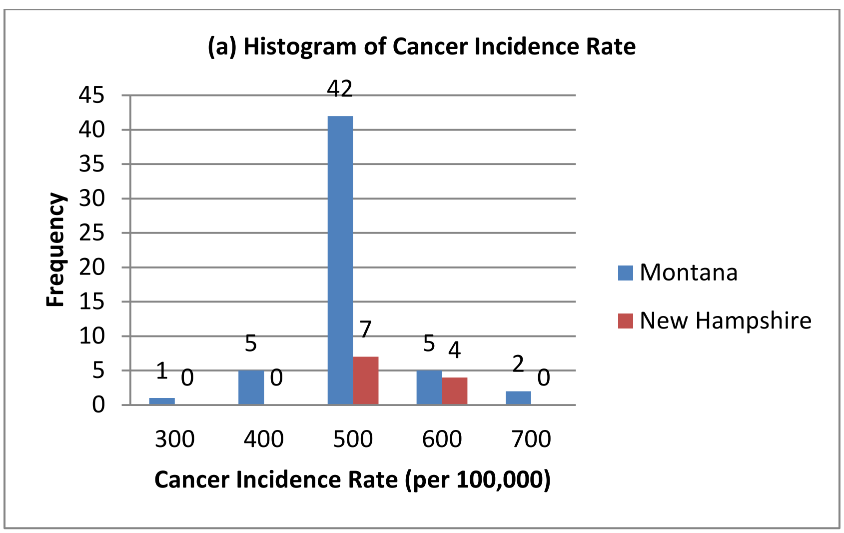
4.1 Cancer data visualization

Cancer incidence rate and death rate in Montana and New Hampshire are presented (Figure 6, 7 and 8). Cancer incidence rates by county in both states are around 500 (per 100,000 standard population). In Montana, Colden Valley, Treasure and Musselshell counties (highlighted in bright blue) have the top three cancer incidence rates, and are adjacently located in central west Montana and , with cancer rates all higher than 550 (Figure 6). The top three highest cancer death rates in Montana are Fallon, Wibaux and Musselshell counties. Fallon County and Wibaux County are next to each other near the east boundary of Montana. Musselshell County would be a very interesting county to investigate in future study, for its low health status measured by both cancer incidence and mortality rates.

From the county level cancer incidence rate map (Figure 7a) and cancer death rate distribution map (Figure 7b) in Montana, among all the public SCC geographic areas, only Fort Peck American Indian reservation area has very high average cancer incidence rate and death rate, while the rest are with middle or low cancer incidence rate areas. It is noticeable that

Broadwater County, in the central west, with very high cancer death rate appears to have a very low cancer incidence rate.

Similarly, the distribution map of cancer data of New Hampshire (Figure 8) shows that cancer incidence and cancer death rate are of high values (in dark brown) in the southeast corner of the state where four public SCC requests in 2013 are outlined in blue. Three adjacent counties of Strafford, Belknap and Rockingham, highlighted in Figure 8, have the highest cancer rates. If we assume that cancer data are even across the whole county, Town of Gilford in Belknap County could be assigned a cancer incidence and death rates of 510.7 and 190.2 rate of 527.8, both of which are very high values in New Hampshire.



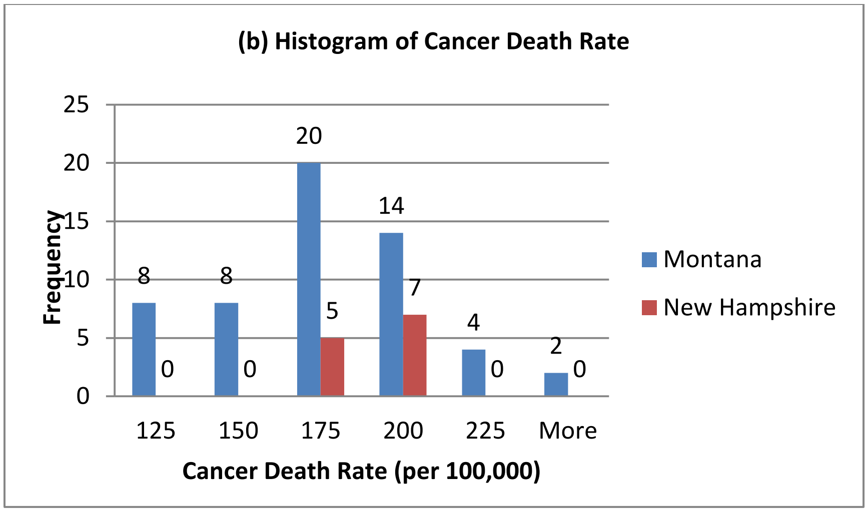
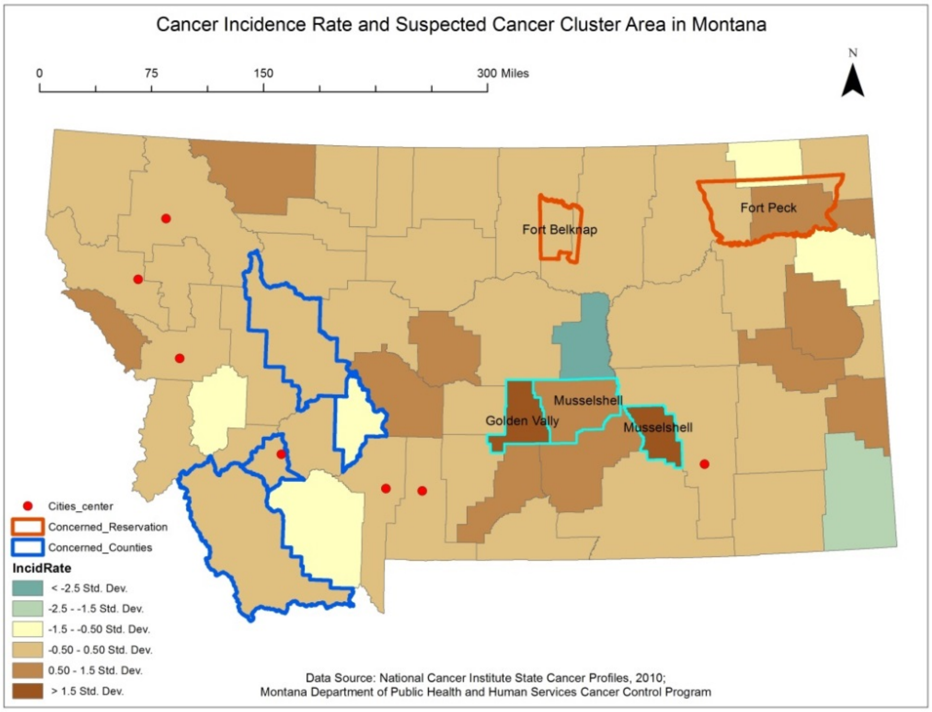


Figure 6 Histograms of Cancer Data in Montana and New Hampshire

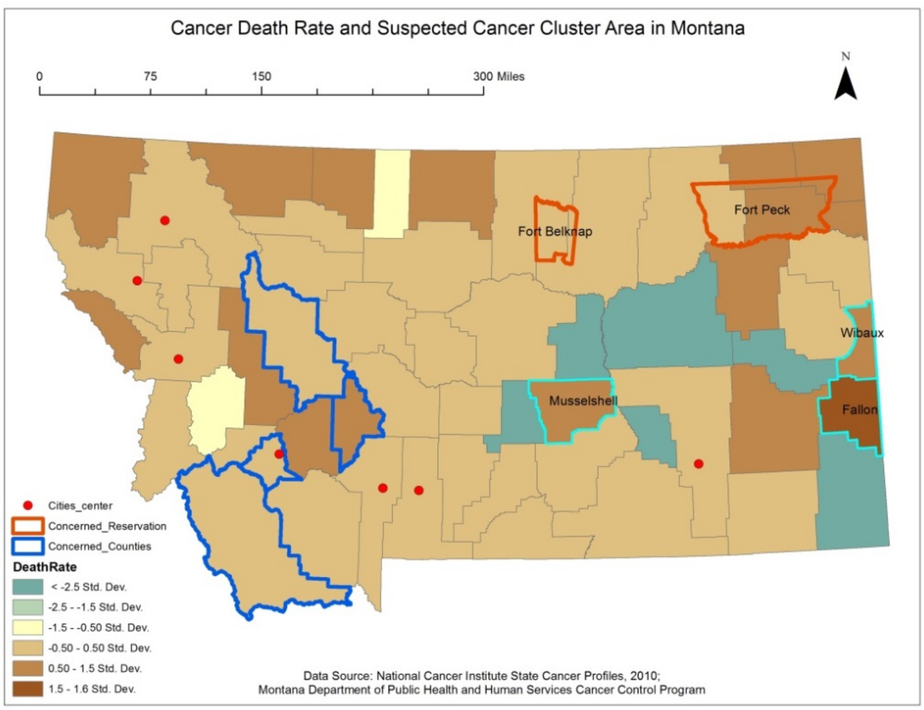
In Delaware, as shown in Figure 9 Kent County has both highest cancer incidence rate (up to 530 per 100,000 people) and death rate (198.8 per 100,000 people). However, public SCC area in Kent County is the smallest.

Table 6 Cancer statistics in Delaware

Delaware County	Incidence Rate	Death Rate
New Castle	491.5	185.9
Kent	532.6	198.8
Sussex	499.3	177.6



(a) Cancer Incidence Rate of Montana



(b) Cancer Death Rate of Montana
Figure 7 Cancer Incidence and Death Rate of Montana

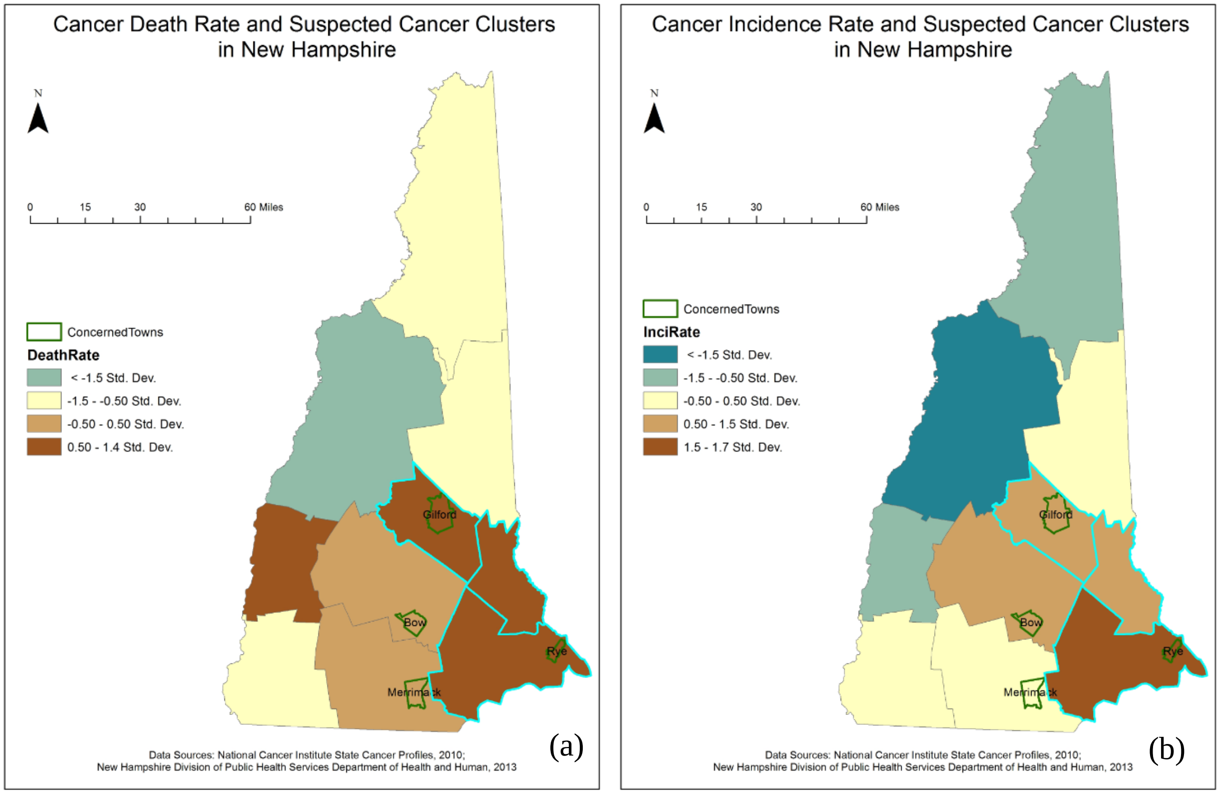


Figure 8 Cancer Incidence and Death Rate of New Hampshire

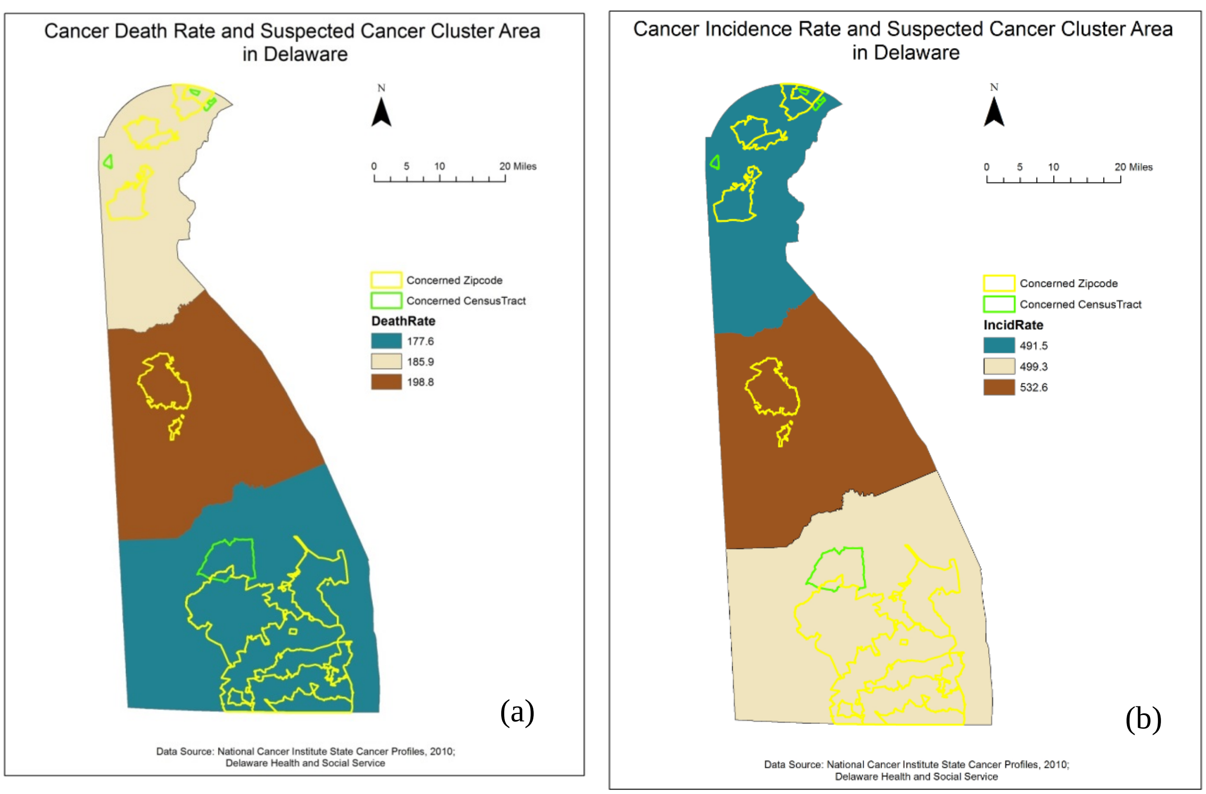


Figure 9 Cancer Incidence and Death Rate of Delaware

4.2 Geospatial and traditional statistical analysis of cancer data

Since several counties in Montana and New Hampshire with high cancer incidence rate and death rate have been identified as “neighboring counties”. Global Moran’s I statistic was calculated on those cancer data. The null hypothesis (H_o) for this test is that the cancer statistics being analyzed is randomly distributed among the counties in study area (ArcGIS Resources 2012). Table 7 summaries the results. There is no Moran’s I’s value significant at 0.05 level. The null hypothesis cannot be rejected; therefore, the two cancer statistics in two study areas were randomly distributed. Additionally, a Getis-Ord Gi* test is performed locally to identify if there are high cancer incidence and death rate close together. The result is displayed in Figure 10. There is a cluster of statistically significant hotspot in the central Montana, but none of them are SCC areas, which reflects the situation mentioned previously that non-SCC areas might be actual cancer cluster areas. For cancer death rate hotspot map in Figure 10, there is no hotspot showing up, instead, 4 cold spot counties are located in the central east Montana, where a SCC city locates.

Table 7 Results of Moran’s I

State	Value	Moran’s Index	z-score	P-value
Montana	Cancer Incidence Rate	0.037162	0.768497	0.442192
	Cancer Death Rate	-0.044054	-0.328864	0.742258
New Hampshire	Cancer Incidence Rate	0.223382	1.954817	0.050605
	Cancer Death Rate	0.010632	0.703436	0.481784

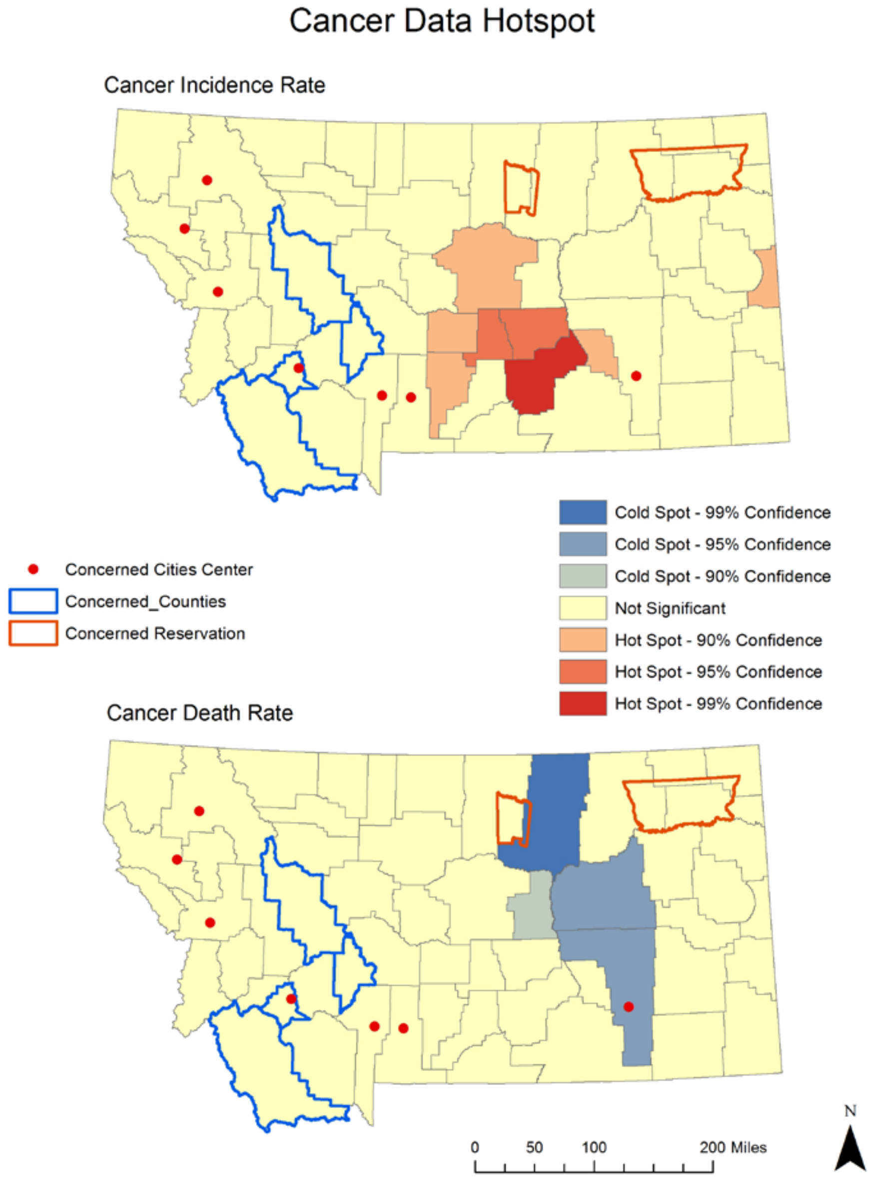


Figure 10 Cancer data hotspot map

In order to compare if cancer data in the public concerned geographic areas is significantly different from the non-concerned areas, student-t test is performed to compare cancer incidence and death rate between two groups whose null hypothesis states that the difference in means is zero. It is problematic when trying to assign an appropriate cancer incidence rate to American Indian Reservation areas (Figure 11) using county level data source.

Areal interpolation was applied to calculate cancer rate in concerned areas in Montanan. The output of t-test (Table 8) indicate that cancer incidence rate of concerned group is slightly smaller than that of the non-concerned area with the rate of 446 and 458 respectively. And the death rate in the concerned group is slightly higher than the non-concerned area's rate. However, since both two-tail p-values are bigger than 0.05, there is no statistically significant difference between concerned areas and non-concerned areas for both cancer incidence and death rate.

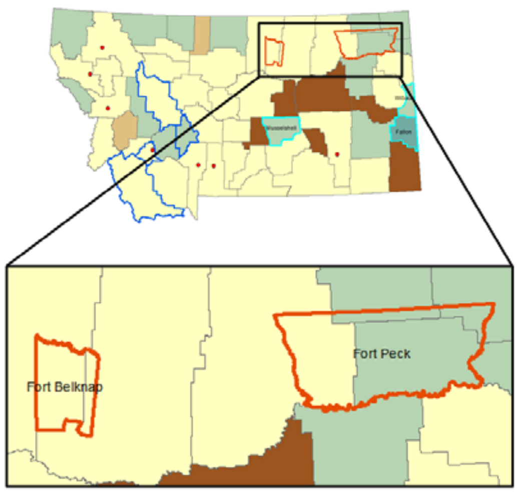


Figure 11 Scale Inconsistent between American Indian Reservation areas and Cancer Data

Table 8 Results of T-test of cancer data from areal interpolation

	Cancer Incidence Rate		*Cancer Death Rate*	
	Concerned	Non-concerned	Concerned	Non-concerned
Mean	446.439	458.065	165.984	153.2725
t Stat	0.89		0.78	
P(T<=t) one-tail	0.19		0.22	
t Critical one-tail	1.68		1.67	
P(T<=t) two-tail	0.38		0.44	
t Critical two-tail	2.02		1.99	

4.3 Geographic distribution of SCC areas and racial group

Black population rate distributions in Montana and Delaware (Figures 12) were mapped by census tract. Firstly, in Montana, few public SCC areas have relatively high ratio of black population, and the overall black population rate in the state is pretty low (0 to 9 percent), while in 2012, the black population made up 14.2% of the total U.S. population (CDC 2014). However, in Delaware, the black rate can be as high as 90 percent in some census tract areas, which largely overweighs that of Montana. Although the concerned zip code and Census Tract did not show high black rate in Delaware either, in northern Delaware, the highest black rate region is encompassed by public SCC areas. In central Delaware, similarly, high black rate areas were right adjacent to public concerned ZIP code area and partially overlapped with those areas. In the following logistic regression, black population ratio is studied in a more quantitative way.

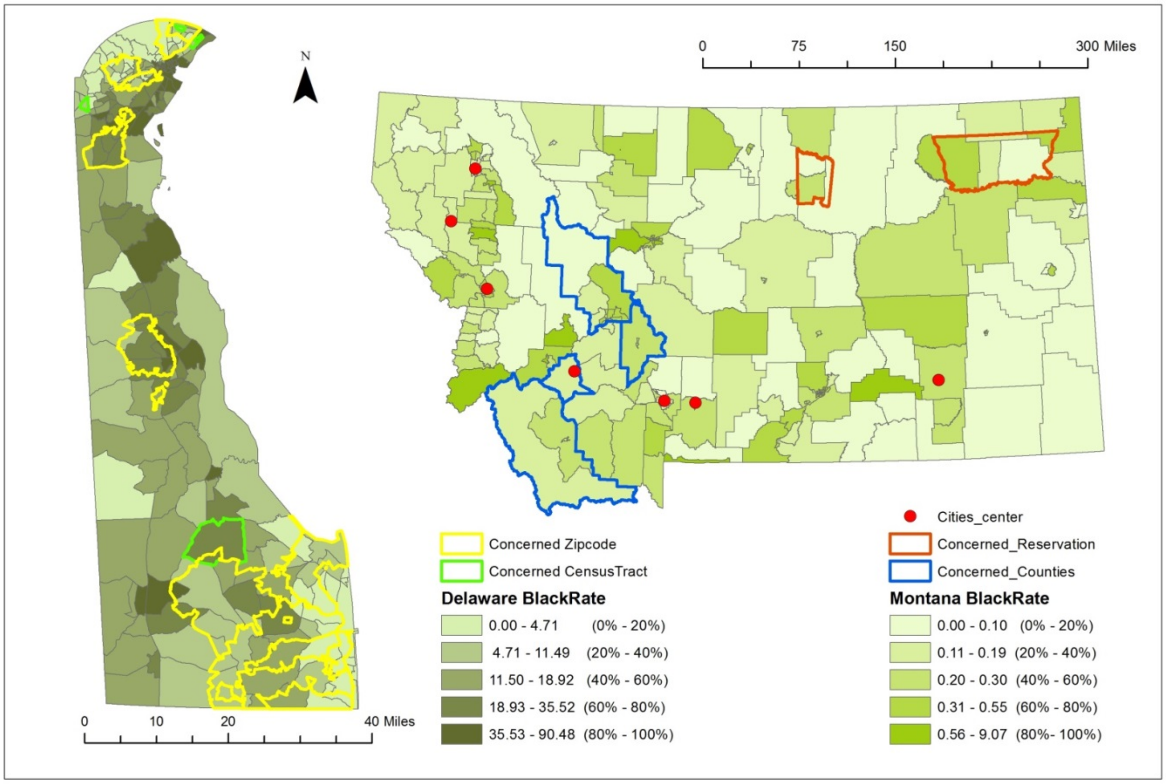


Figure 12 Black Rate and SCC Areas in Montana and Delaware

4.4 Geographic distribution of SCC areas and age group

The distribution of population in elderly age group has also been analyzed in Montana and Delaware (Figure 13). In Montana, nearly half of concerned areas belong to the highest quintile section of >65 years old ratio (In dark green). Two American Indian Reservation areas are dominated by Census Tracts with low elderly population ratio (light green) which range from 5 to 16 percent. However, cancer incidence and death rate in Fort Peck Reservation were high, indicating that there might be more middle aged or young people that have being diagnosed with cancer in this area. In Delaware, the top 15 Census Tracts with highest elderly population ratio are clustered in southeast corner (Figure 14), which is largely overlapped with concerned ZIP code. Looking northwards, also, there is a Census Tract cluster with high ratio of elderly population, in addition to the public SCC geographic areas being covered by a portion of high ratio areas. And the concerned area in central Delaware is close to Census Tract with high elderly population ratio. Therefore, in Delaware, there might be some association between age-group ratios and public request for SCC investigations.

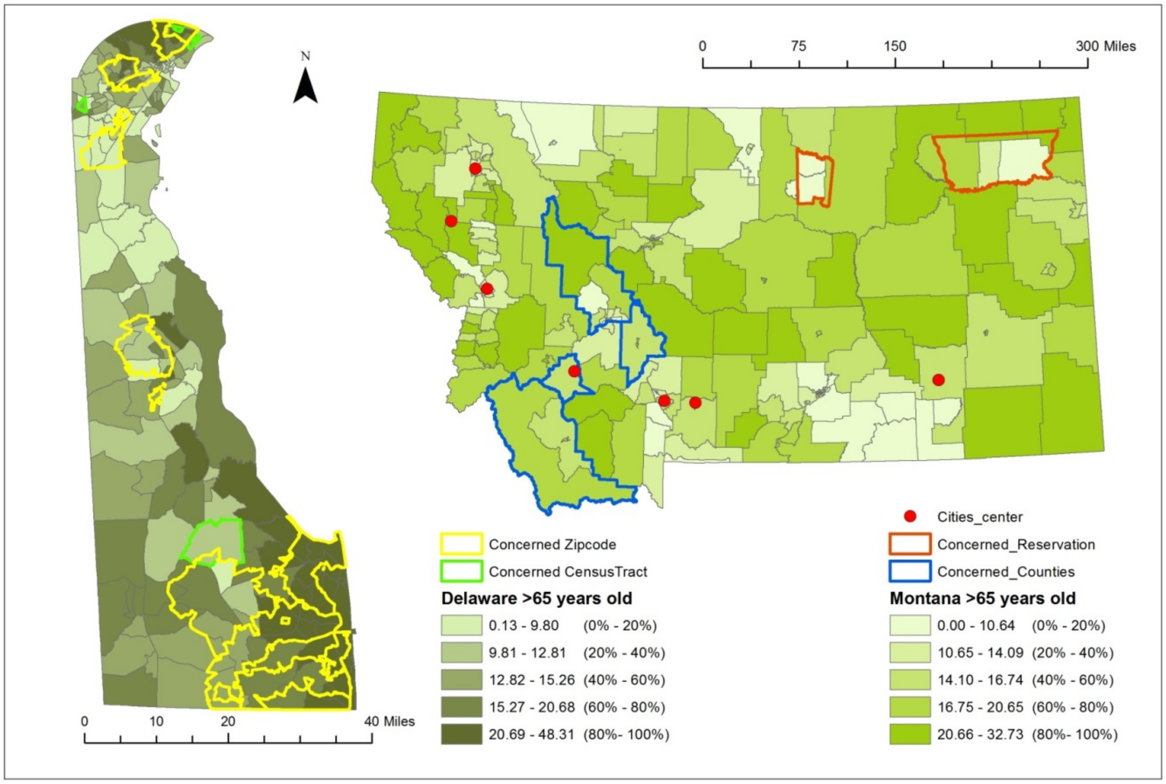


Figure 13 Elderly Age Group (>65 years old) and SCC areas in Montana and Delaware

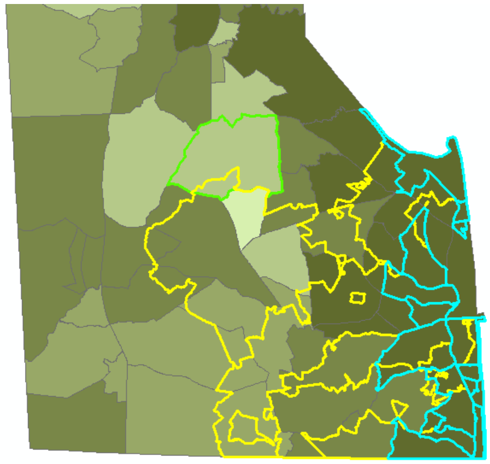


Figure 14 South Delaware (Top 15 Census Tracts with highest elderly population ratio)

4.5 Geographic distribution of SCC areas and income and education

Environmental related studies demonstrate that economically disadvantaged communities are often disproportionately located in close proximity to industrial sources of pollution, and often these pollutants are known to cause cancer (Jordan, 2012). This study examines the distribution of median household income and education level at the Census Tract level.

Take a look at the distribution maps of two SES variables in Figure 15 and16, many high household income Census Tracts located in the south Montana. Two American Indian Reservation Areas are very low income area, as well as less education area. Overall, the Census Tracts with high ratio of bachelor degree population are located in the southwest Montana. Several regions with both high SES variables could be identified in the maps.

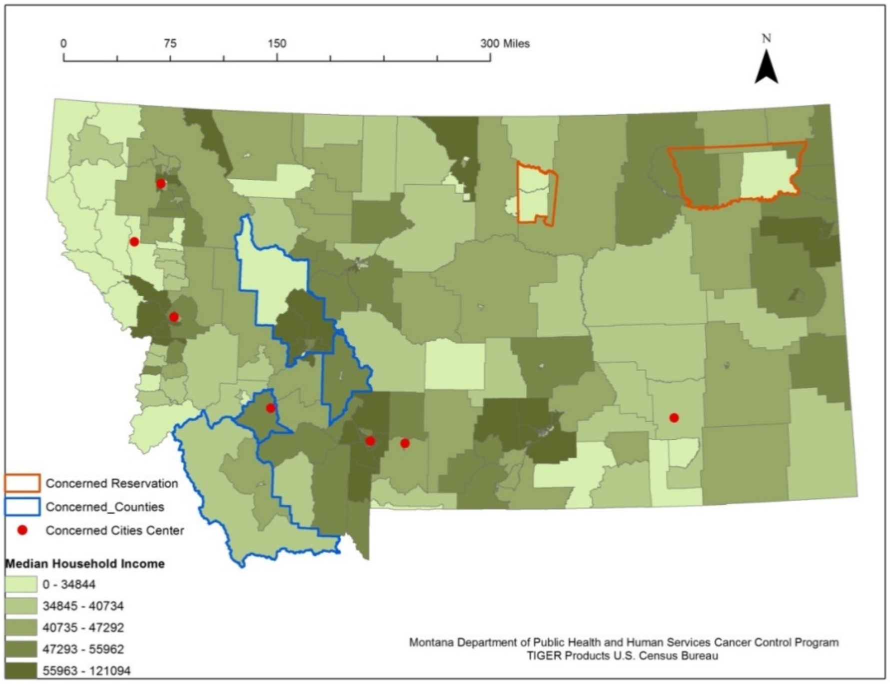


Figure 15 Distribution map of Median Household Income

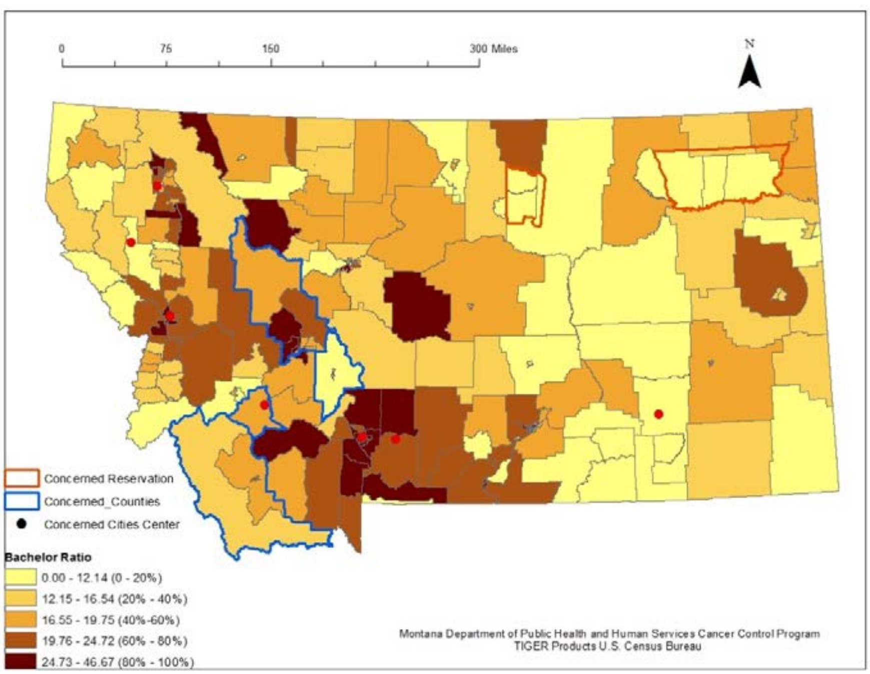


Figure 16 Distribution Map of Bachelor Ratio (among population 25 years and over)

4.6 Environmental health assessment for EPA data

In terms of environment contaminant implication, the distribution of TRI sites in Montana was used as an example to explore further. Figure 17 is a Euclidean distance surface from every cell to the nearest TRI sites. Almost all SCC geographic areas have at least one pretty close TRI sites which is within 50 kilometers (in red). However, only Fort Belknap Reservation area did not have a nearby TRI site closer than 50 kilometers, and its distance to the nearest TRI site is more than 200 kilometers (in green).

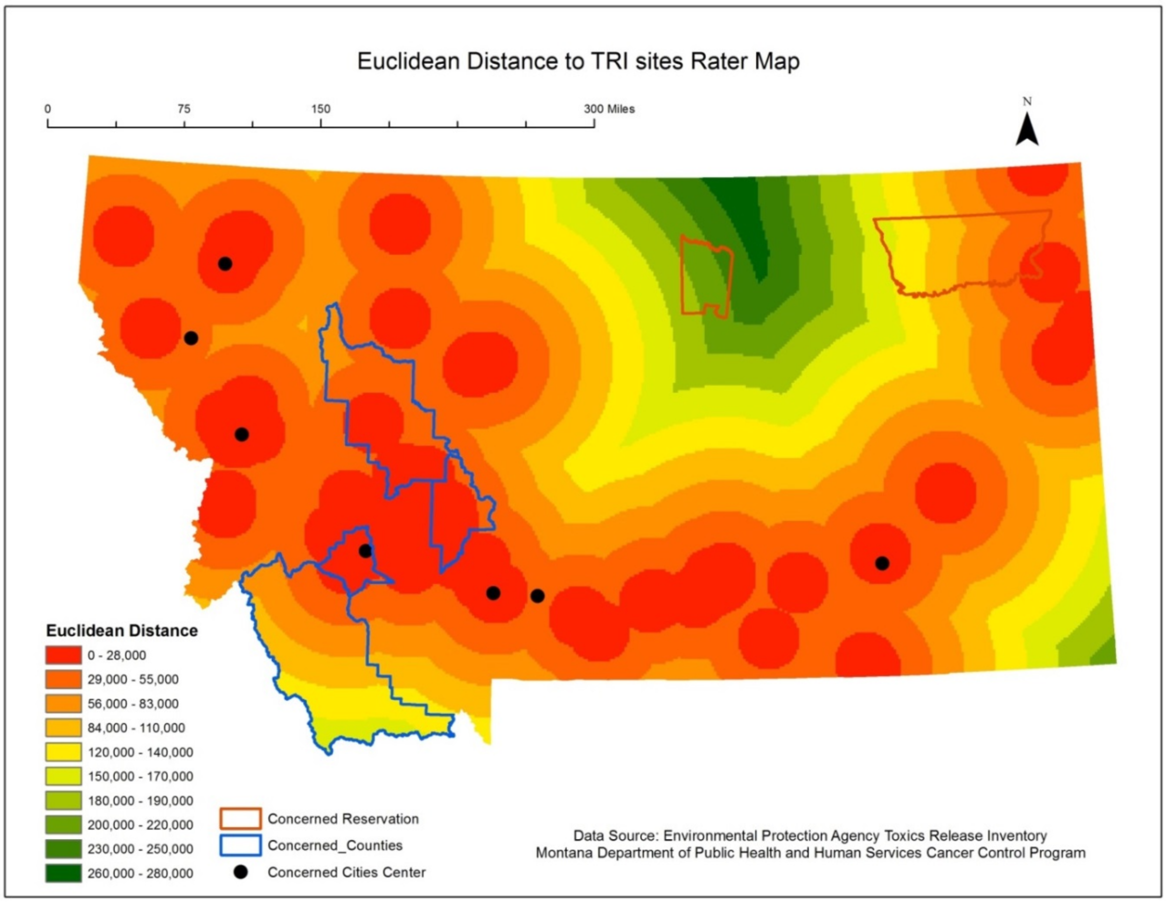


Figure 17 Euclidean Distance to TRI sites in Montana

Transparent blue areas on Figure 18 are 5-mile buffer areas of public concerned SCC areas. The red points represent TRI sites. TRI facilities are mainly located in central west and south Montana. For city of Missoula, there are 8 TRI facilities. Silver Bow County, Lewis and

Clark County, and Broadwater County form a small cluster of geographic areas with large number of TRI sites. Four SCC areas do not have any TRI facility inside or within 5 miles: the two Indian American Reservation areas, City of Hot Springs and City of Livingston (labeled in green). Moreover, 1-mile buffers on TRI sites are also calculated. There are a total of 54 out of 215 square miles of buffered areas intersected with concerned geographic areas, approximately 25 percent.

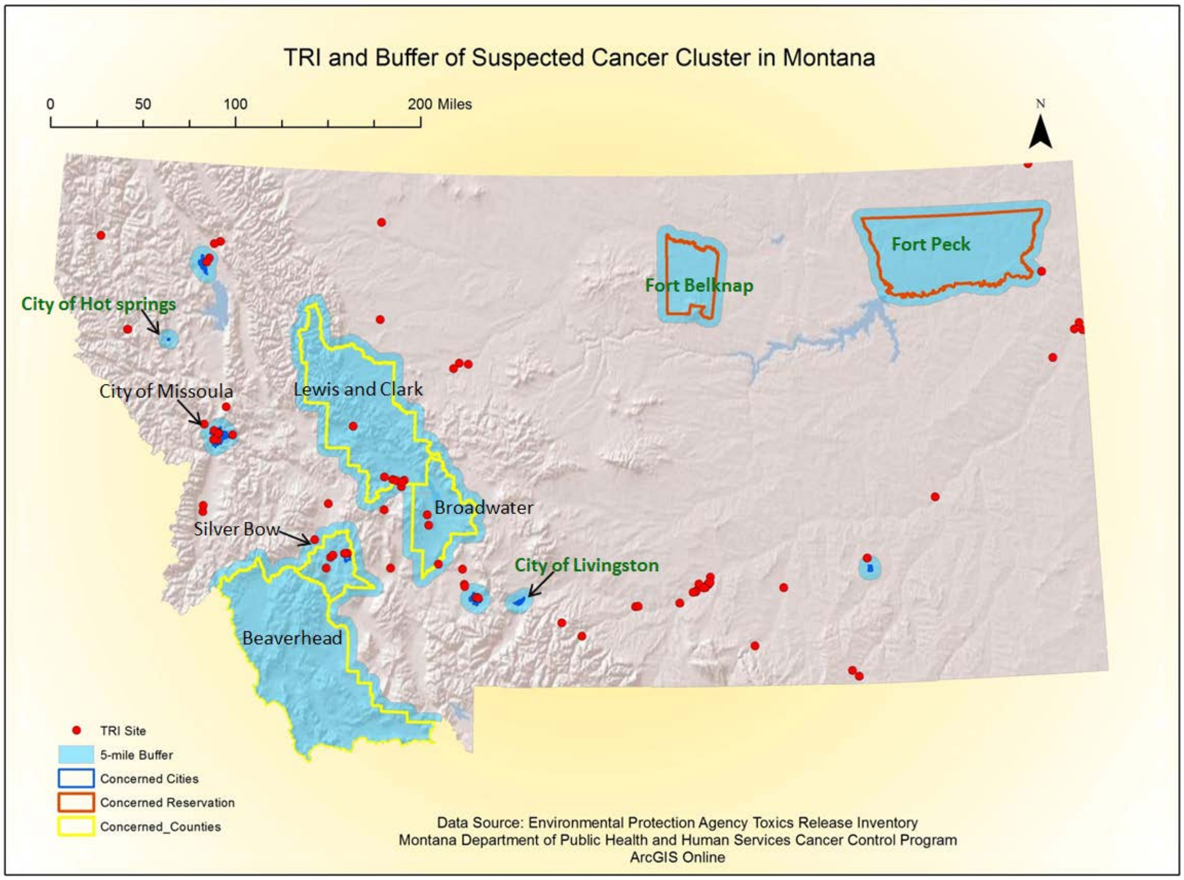


Figure 18 TRI and 5-miles Buffer of SCC in Montana

The distribution of impaired waters and its relationship with TRI sites and concerned geographic areas can be seen in Figure 19. Impaired waters are concentrated in southwestern Montana, which is crudely matched with the distribution pattern of TRI sites and concerned geographic areas. Every concerned geographic area somehow has impaired water inside or on its boundary. Additionally, it is obvious that the majority of TRI sites are located either right along or fairly close to impaired waters, which indicates that TRI sites probably bring negative impact on surrounding water quality.

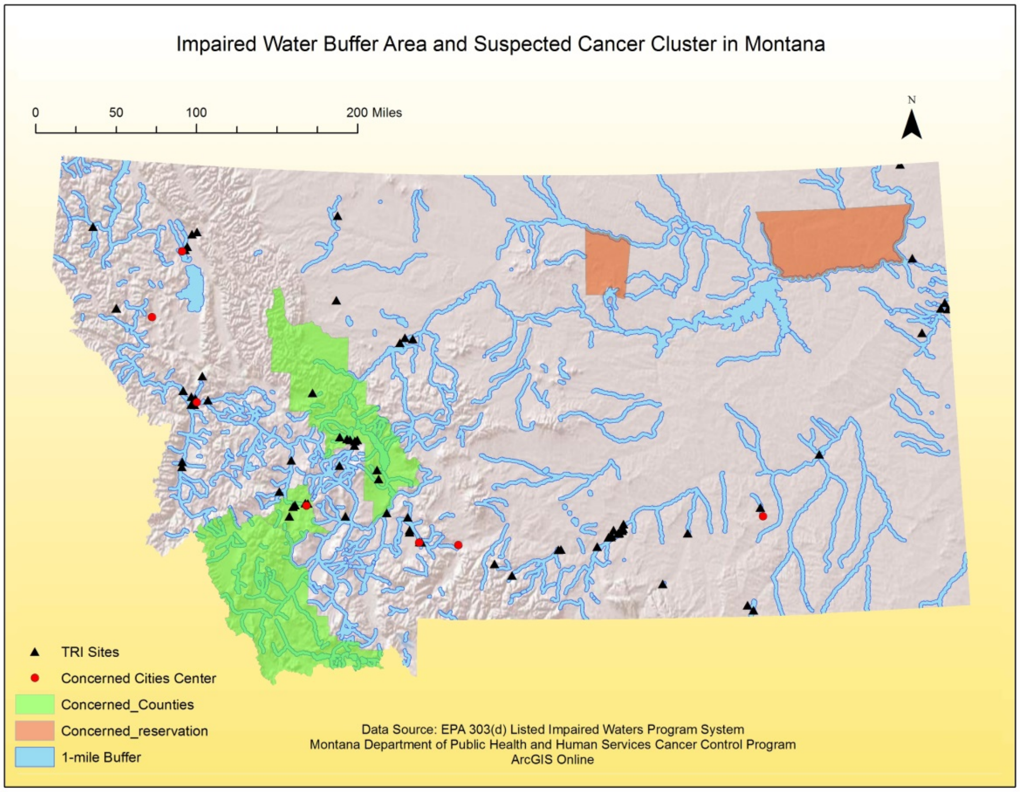


Figure 19 Impaired Water Buffer Areas in Montana

4.7 NURE interpolation mapping

Figure 20 displays the predict concentration maps of the selected three elements from NURE sediment dataset: uranium, lead and zinc. Not every chemical concentration has been analyzed, because, on one hand, many chemicals' sample sites did not cover the entire study area, so the number of points is big not enough to perform spatial interpolation. On the other hand, the selected elements are harmful ones. For uranium, intakes of exceeding amount and lead to increased cancer risk or lead to internal irradiation and/or chemical toxicity (EPA). Expose to uranium can lead to DNA damage and result in genetic mutations, chromosomal aberrations which as facilitate carcinogenesis (Wagner 2011). And the high lead values being mostly of anthropogenic origin (Lima 2003), which might indicate that hazards industries with emission of lead are nearby. And also the synthetic assessment map (Combine) which classifies and labels the combined concentration from 1 to10. Firstly, the upper two maps in Figure 20 have similar pattern. The entire west region is dominated by high concentrations where many concerned SCC areas are located. Especially for the region between concerned counties (blue polygon), it is noticeable that a large area of extremely high uranium concentration is shown. Compare these two maps with Figure 19 (impaired water), the distributions are very similar, because the impaired water in concerned counties and the nearby area are very clustered. For the predicted lead map, we can see multiple high value spots are clustered in concerned counties. Recall figure 19 again, high lead concentration spots are good match with the location of several TRI sites clusters, which confirms the assumption that the high lead values being mostly of anthropogenic origin (Lima 2003), which might be caused by the nearby hazards industries that release lead or its compound.

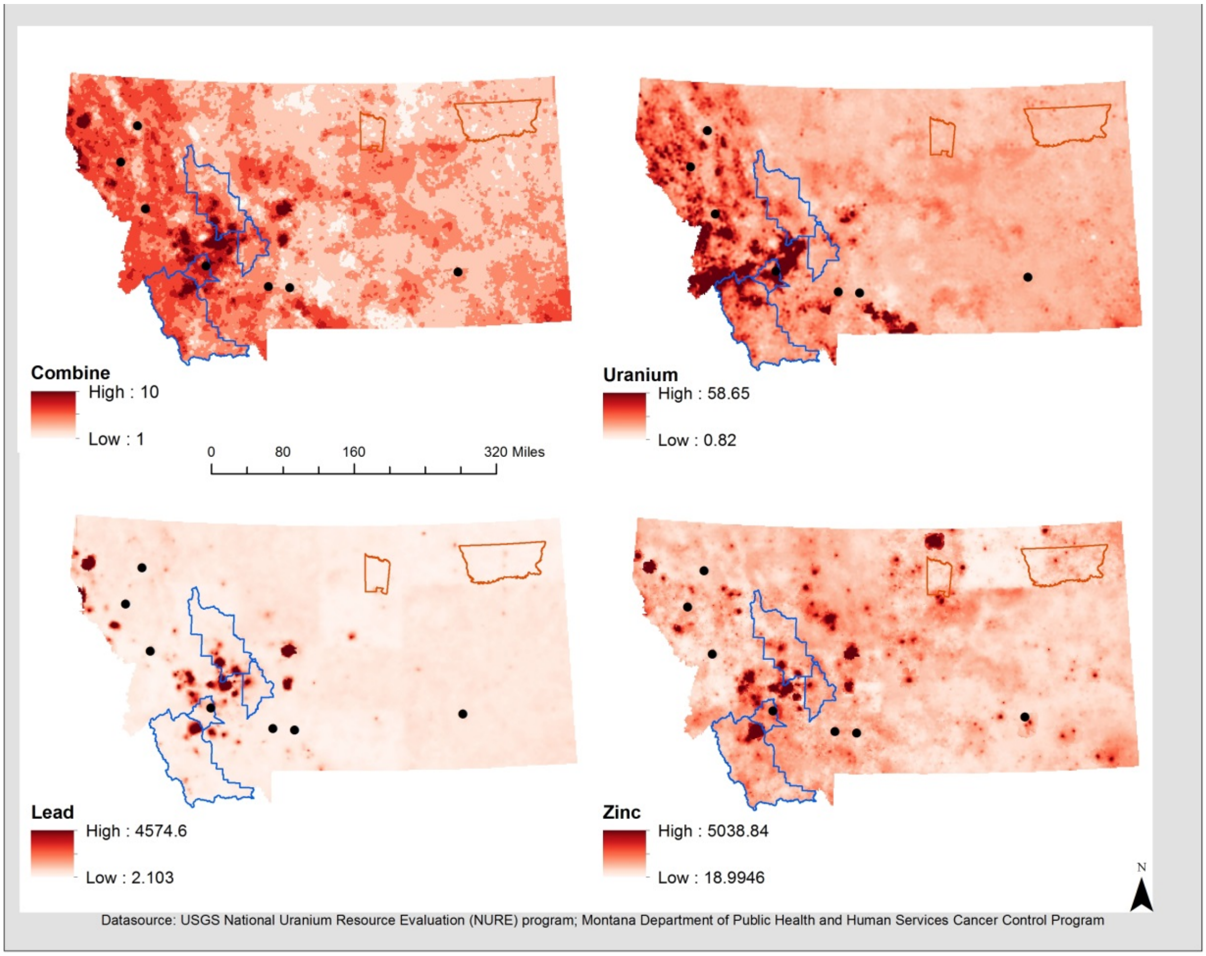


Figure 20 Predicted chemical concentration

4.8 The logistic regression

Grouping the Census Tract into concerned and non-concerned group is the basis of further statistical analysis. Therefore, Figure 21 demonstrates two Census Tract groups in coral red and beige. It shows that concerned Census Tracts (red) are good surrogates for public concerned areas. The concerned counties (blue) are perfectly matched with several Census Tracts, and for American Reservation areas (orange), the difference are minor. Therefore, it is generally reasonable using Census Tracts to represent concerned and non-concerned areas in the following studies.

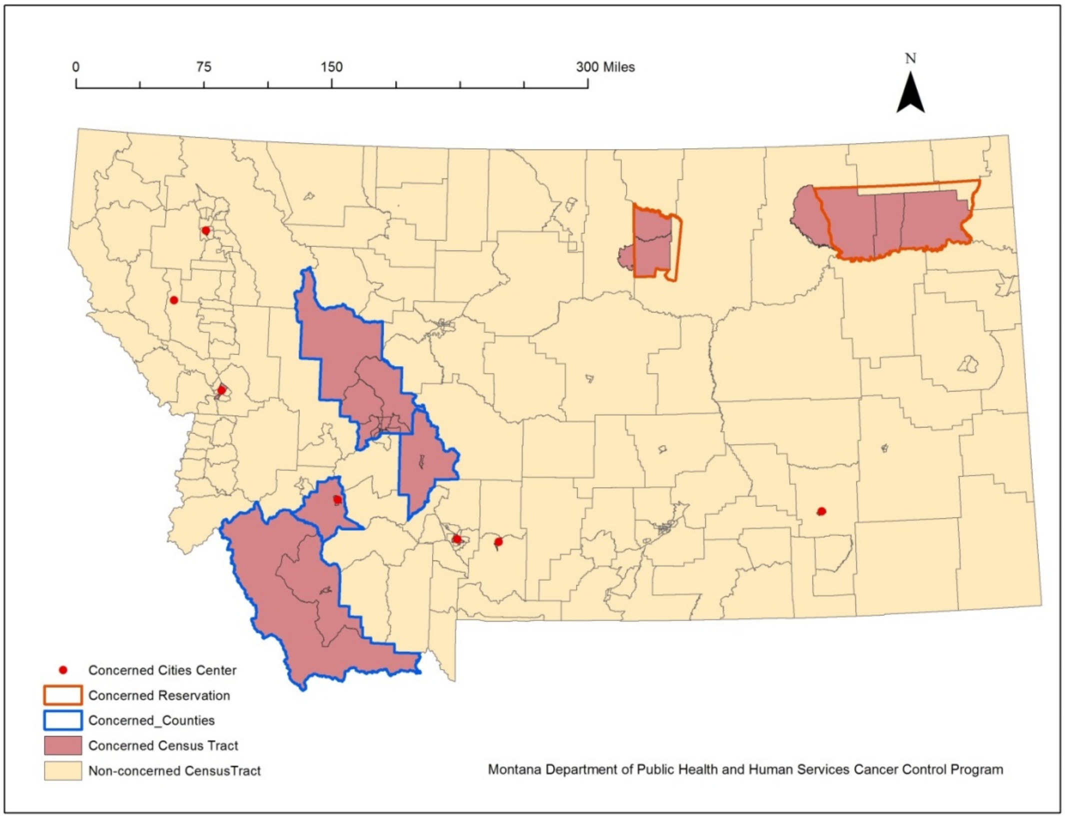


Figure 21 Concerned/Non-concerned Census Tract Group

With two groups of Census Tract, the nature of variables for the logistic regression is analyzed, based on the assumption that residents in concerned geographic areas do have bad health outcome and this is derived from environmental health injustice. Table 9 provides descriptive statistics of the variables, which are derived from the collected dataset. Mean values for concerned and non-concerned Census Tract groups are listed separately, t-test is also performed to test the difference between groups which is helpful to see if public SCC reveals any disparities. Firstly, speaking of cancer data, average cancer death rate in concern group is 162.04, and that of non-concern group is 164.44. The difference is only 2 out of 100,000 and not statistically significant. While, for cancer incidence rate using Census Tract as the basic unit, the difference between concern and non-concern groups is statistically significant, with p-value smaller than 0.001. That is to say that, averagely, cancer incidence rate in SCC areas is lower than the rest, which is on contrary to my expectation.

For socioeconomic and demographic variables, the difference of population density between concerned group and non-concerned group is statistical significant. Mean population density in concerned Census Tracts is about 3 times of the population density in non-concerned Census Tracts, which is align with our common sense that SCC reports are more likely to occur merely because of high population density. The mean value of bachelor education and percentage of elderly (>60 years old) population in concerned and non-concerned groups are significant with both p-values smaller than 0.001. In public concerned Census Tract group, it tends to have less elderly people, but more people owning a bachelor degree. The mean value of median income and percentage of black people of concerned group are lower than these of non-concerned group, but the disparities are not significant.

For environmental variables, TRI total releases per square kilometer in concerned geographic statistically significantly overweighs non-concerned areas, with a p-value of 0.02 and mean values of 20282 tons/km^2 and 1133.32 tons/km^2 respectively, which is consistent with the visual analyses on maps in the previous section. Based on interpolation results and zonal statistics, the disparities of mean of uranium concentration and lead concentration across concerned and non-concerned groups are significant at a 0.001 level. For chemical uranium, average concentration of concerned group is 4.85 ppm, which is larger than non-concerned group of 3.73 ppm. Mean lead concentration is concerned group is more than two time of non-concerned group. As for the statistics of zinc and maximum concentrations in a Census Tract, none of them has significant disparity.

Lastly, both the mean and maximum value of combined chemical concentrations in concerned group significantly overweighs the rest Census Tracts at a 0.001 level. As well as the reclassified lead concentration becomes significant, and the reclassified grades of mean lead concentration in concerned and non-concerned groups are 3.69 and 2.53 respectively. Similar to the original mean uranium concentration, reclassified uranium of concerned group is 1 grade higher than non-concerned group. However, neither of the reclassified value and original value for zinc shows significant difference.

Table 9 Descriptive statistic and t-test: Census Tracts by SCC reports

	Mean of Census Tracts		**T-test (p-value)**
	Concern	Non-concern	
Cancer variable			
Cancer Death Rate	164.44	162.04	0.34
Cancer Incidence Rate ***	**442.75**	**460.19**	**0.00019**
Socioeconomic and Demographic			
Population Density(per km^2) ***	726.42	276.60	<0.0001
% Bachelor Degree ***	**22.15**	**18.12**	**0.0003**
Median Income	44852.62	46692.64	0.38
% elderly population ***	**12.94**	**16.10**	**0.0001**
% black population	0.36	0.41	0.60
Environment			
TRI release (tons/km^2) **	**20282.24**	**1133.32**	**0.02**
TRI total on-site release	2165469.43	546129.10	0.12
Air quality nonattainment status **	**0.64**	**0.48**	**0.03**
Maximum uranium concentration	6.92	7.19	0.79
Mean uranium concentration ***	**4.85**	**3.73**	**0.0002**
Maximum lead concentration	156.93	125.63	0.63
Mean lead concentration ***	**51.60**	**22.68**	**<0.0001**
Maximum zinc concentration	204.96	271.73	0.33
Mean zinc concentration	92.85	94.58	0.87
Reclassified chemical concentration			
Maximum combined chemical concentration (from 1-10) ***	**4.81**	**4.24**	**0.009**
Mean reclassified combined chemical concentration ***	3.81	3.05	<0.0001
Mean uranium concentration (reclassified) ***	6.755.73	5.73	0.00018
Maximum uranium concentration (reclassified)	8.17	7.79	0.18
Maximum lead concentration (reclassified) ***	5.14	3.85	0.00013
Mean lead concentration (reclassified) ***	3.69	2.53	<0.0001
Maximum zinc concentration (reclassified)	2.91	3.05	0.52
Mean zinc concentration (reclassified)	2.10	1.96	0.21

* p-value < 0.1.
** p-value < 0.05.
*** p-value < 0.001.

Beyond comparing mean values between concerned and non-concerned groups, correlation analysis among variables has also conducted. The results are visualized in Figure 22. The legend bar on the right indicate the coefficient of paired variables with blue and red representing positive and negative respectively. And the ellipses in the grids indicate the correlation results. However, the grids with numbers filled in indicate insignificant correlations, and the numbers are p-values which are all above 0.05. From the first row, the presence of SCC and cancer incidence rate and elderly population are both negatively related with ellipses shown in light orange. Whilst, it shows positive correlation with higher bachelor rate, more TRI releases more uranium and lead in sediment, as well as, bad air quality. These trends are consistent with the previously mentioned t-test. For the correlation results between independent variables, cancer death rate and incident rate are positively related. Similarly, the higher the death rate is the higher the lead concentration, and the worse the air quality it. High cancer Incidence rate is also associated with bad air quality. For SES variables, bachelor rate is positively correlated with income, but negatively correlated with elderly population. Black rate is negatively related with income, but positively related with air pollution. In the lower right corner of this correlation matrix, many positive correlations can be identified between various combinations of environmental variables.

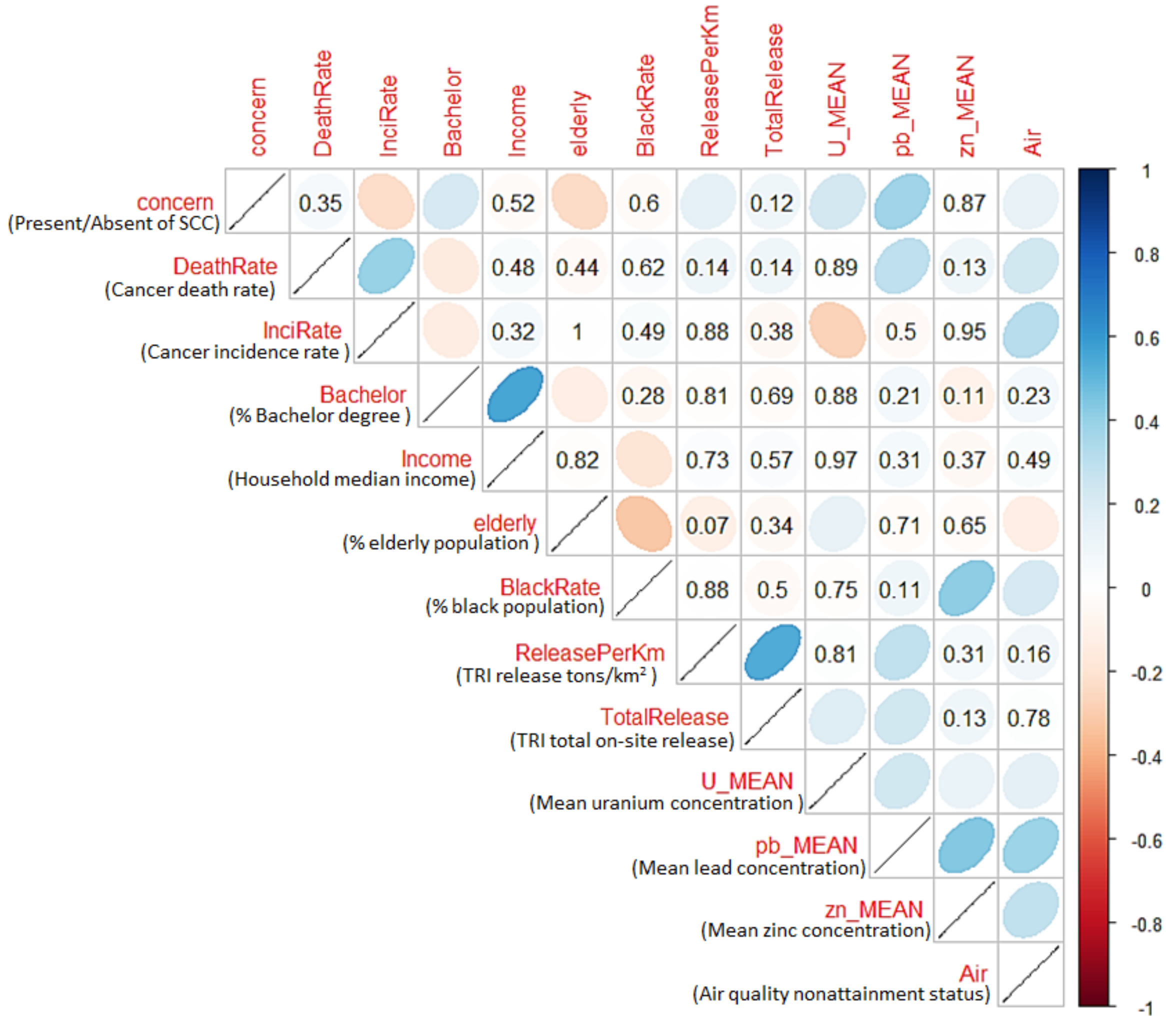


Figure 22 Correlation matrix

With aforementioned variables, in this study, 3 logistic models are built: Environmental model using all the variables, using environmental factors only, SES model using SES factors and cancer data and STEP model using a subset of all variables which is generated from a R built-in function “step()” . This function choose model by AIC a stepwise algorithm, select the variables by AIC (James 2013). For ordinary least square (OLS) regression, R-square indicate the proportion of variance explained by the predictors, however, logistic regression does not have the equivalent parameter. Therefore, AIC could be one of the suitable criteria for selecting variables. When working with many input variables, multicollinearity is often concerned. There are a number of measures of multicollinearity, variance inflation factor (VIF) is used to test for multicollinearity for logit model in R. A VIF measures how much the variance of an estimated coefficient is increased because of collinearity. All of the VIFs of variables in the model is smaller than 3, so there is no multicollinearity problem with the variables, since, as a general rule, multicollinearity is assumed to be high when VIF is larger than 5 (Psychstatistics 2012).

Table 10 shows the predicted odds ratio for two models separated by environmental and SES factors, together with their statistical significance. The column odds ratio is calculated by exponentiating the predicted coefficients, so that we can interpret them as odds-ratios instead of log odds. Now we can say that a one unit increase in lead mean concentration, the odds of being reported as a SCC area versus not being reported increase by a factor of 1.573. In SES model, the median income with an odds ratio of 1 does not change the odds of being reported as a SCC area. Bachelor ratio actually plays a good role in increasing the odds of being reported as a SCC area. The odds ratio that belongs to the variable "%elderly population" negatively departure from 1, and let to the inference that the elder population ratio has an effect of lowering the odds of being reported as a SCC area. The odds ratio of parameter "Cancer Incidence Rate" and "Cancer Death Rate" is close to 1, which indicates that it has a relatively little impact on the occurrence of SCC reports. However, the corresponding p-values are extremely small, indicating that (1) with cancer death rate increases, the SCC report is more likely to occur, (2) the decrease of cancer incidence rate tend to increase the likelyhood of SCC report occurrence at a Census Tract scale.

Table 10 The regression results obtained for environmental model and SES model

Independent Variable	Odds Ratio	p-value
Environmental Model		
combined chemical concentration	0.018	0.42
Mean uranium concentration	0.371	0.29
Mean lead concentration	1.573	0.00237 **
Mean zinc concentration	4.296	0.46
Air quality nonattainment status	0.671	0.28

TRI release (tons/km^2)	0.638	0.88
SES Model		
Median Income	1	0.0019 **
% Bachelor Degree	1.125	<0.001***
% black population	0.412	0.082
% elderly population	0.892	<0.001***
Cancer Incidence Rate	0.975	<0.001***
Cancer Death Rate	1.048	<0.001***

* p-value < 0.1.
** p-value < 0.05.
*** p-value < 0.001.

Table 11 presents the estimated odds ratio for the STEP model, together, also, with their statistical significance. The optimized combination of independent variable includes population density, cancer incidence rate, black population rate, elderly population rate and uranium, zinc and lead concentration. The odds ratios that belong to variables “Population density”, “Lead” and “Uranium” are bigger than 1 leading to the inference that the population density and mean lead and uranium concentration in a Census Tract have a positive relationship with the increase of odds of SCC reports presence. And among these two, lead and uranium concentration plays an important role with a higher odds ratio, indicating that a one unit increase in lead or uranium mean concentration, the odds of being reported as a SCC area versus not being reported increase by a corresponding factor. Odds ratio belong to population density is too close to 1, indicating that increase in population density will not lead to a large increase in the odds of SCC report presence. However, odds ratios of variable “Cancer Incidence Rate”, “% black population” , “% elderly population” and “Zinc” are smaller than 1, among which black population ratio plays an important role with a odds ratio of 0.12 which strongly depart from 1. Basically, the increase of cancer incidence rate, black population rate and elderly population rate will potentially bring the effect of decreases of odds of SCC report occurrence.

Table 11 The regression results of STEP Model

Independent Variable	Odds Ratio	p-value
Population density	1.002	5.76e-06 ***
Uranium	1.138	0.110074
Lead	1.037	2.44e-06 ***
Zinc	0.989	0.034417 *
% black population	0.120	0.016279 *
% elderly population	0.876	0.000762 ***
Cancer incidence rate	0.975	0.000376 ***

* p-value < 0.1.
** p-value < 0.05.
*** p-value < 0.001.

4.9 ArcGIS Online web application

As a broad impact of this thesis research, an interactive web map application displaying public suspected cancer cluster areas in Montana has been created and published online that is accessible to everyone (http://tinyurl.com/ppzkp6v). The web page, titled Public Suspected Cancer Cluster Investigation, is shown in Figure 21. Public SCC area and concerned census tract area in Montana are displayed in transparent polygons in different color respectively. The tool box in the top left allows user bring up legend, base maps, layers, including TRI sites, cancer incidence and death rate. TRI sites are displayed in circles whose sizes are proportional to the accumulated toxic chemicals they released. Cancer data are displayed based on quantile classification scheme. The search box facilitates users zoom into the address they care about. Additionally, the interactive popup window of each feature provides more detailed information. For instance, in Figure 21, the popup of a TRI sites tells user the name, industry, location, and accumulated onsite releases of this TRI site. Users are able to examine the geospatial relationship between these features and visualize and explore the spatial distribution.

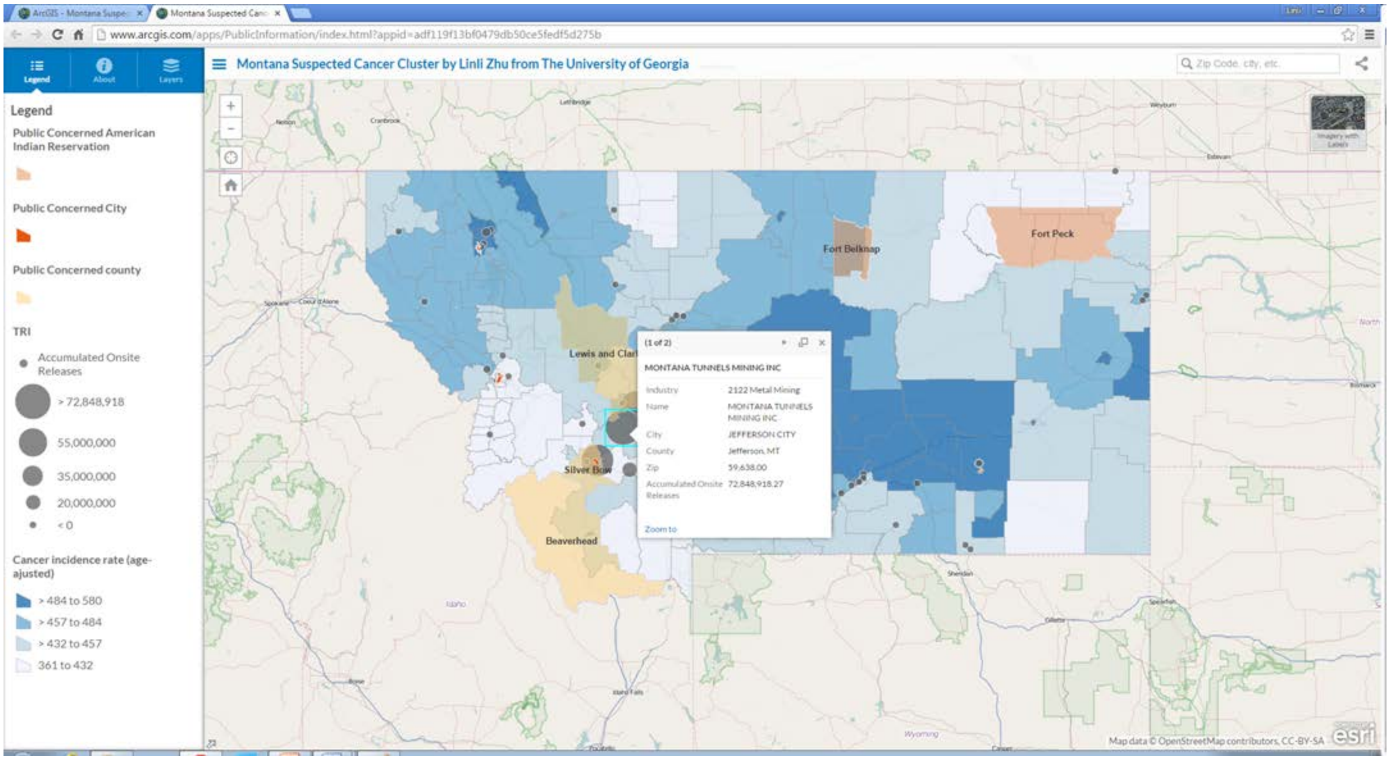


Figure 21 Public suspected cancer cluster web application

In sum, the key results are:

- The majority of SCC areas are urban areas with high population density.
- In public SCC areas, it tends to have less elderly people, but more people owning a bachelor degree.
- The increase of black population rate and elderly population rate will potentially bring the effect of decreases of odds of SCC report occurrence.
- With cancer death rate increases, the SCC report is more likely to occur. However with cancer incidence rate increases, the SCC report is less likely to occur.
- TRI accumulated releases per square kilometer in concerned geographic overweighs non-concerned areas.

- Average uranium and lead concentration in concerned geographic areas are larger than that of non-concerned areas. And lead and uranium concentration plays an important role in increasing the likelihood of presence of SCC.
- The interactive web map application broads the impact of this thesis.

CHAPTER 5

CONCLUSION

This research has compiled public SCC reports dataset in three states in the U.S, explored characteristics of SCC areas by integrating the top-down authoritative cancer data, environmental data and social economic data. Additionally, a good information platform that shares public SCC related information has been developed.

By surveying over twenty state health departments, the scarcity of public accessible SCC data has been uncovered. The data of the concerned geographic areas has not been systematically collected and organized in the U.S or even at the state level. Recently, Department of State Health Services in Texas also published the cancer cluster investigations from 2010 to 2014, which could be potentially a new data source for SCC analysis (Cancer Cluster Investigations 2015). The majority of SCC areas are urban areas along major road. And the geographic scale reported by concerned individual varies either from or within states. In Montana and Delaware, the scale varies dramatically, but in New Hampshire, reports tend to be uniform. Therefore, in future study, how to deal with various reported scale could be one aspect to think about. We may propose different analytical methods when facing different size of a SCC region. Or, we may provide guidelines or advisement when an individual own cancer cluster concerns.

Analyses and results show a range of significant correlates of presence and absence of SCC reports. Firstly, linking the top-down authoritative cancer data to bottom-up SCC requests data, the results of statistical analyses of Montana based on spatial interpolation methods do not show strong disparities in the public SCC geographic areas. However, when I use unified scale

(Census Tract) to represent SCC areas, the result of statistic test becomes significant, but slightly points to the opposite direction of the original intuitive expectation, since the mean value of cancer incidence rate in concerned Census Tracts was significantly lower than non-concerned Census Tracts. One of the possible explanations of this disparity would be the cancer screen facilities are relatively worse equipped in some SCC geographic area leading that some people is not diagnosed effectively but they actually got cancer. Additionally, using County level derived Census Tract level cancer incidence rate data as an input of logistic regression, the results always show negative correlation between incidence rate and odd of SCC requests presence.

Secondly, the results of analyses of demographic and socioeconomic characteristics, including racial, age-group, household income and education level, show heterogeneous spatial relationships either between different states or within one state. The distribution of black population tends to be positively related with SCC areas in Delaware, and similarly for elderly population. High ratio of elderly population and black population areas tend to appear around SCC areas. But no significant pattern is visually identified in Montana. While, thanks to the t-test on socioeconomic variables between concerned Census Tracts and non-concerned Census Tracts in Montana, some disparities in two groups are revealed: (1) Residents in SCC areas tend to be better educated; (2) ratio of elderly people tend to be lower in SCC areas. Recall that the coefficient of elder ratio in logistic model can also be the evidences of the disparities in age-group. Also low black population ratio turns predict higher probability of SCC requests occurrence. In light of those results, I propose that risk perception level might vary from age-groups, racial groups, economic statuses, and more. Well educated people and white people might tend to pay more attention on community health status. Young people might be more informed about the mechanism of SCC reporting methods and system, or more informed about information within community, since young people have more opportunities to communicating with others not only via physical social network but also virtual social network. Additionally, health conscientiousness also varies from individual to individual. Conscientiousness is one of the important aspects of health-relevant personality (Friedman 2000). Some studies show conscientiousness has positive association with SES predicting social environmental factors in terms of education, career success and earnings, and so on (Bogg 2013). Conscientiousness has consistently shown positive association with longevity and negative association with morality and many diseases (Bogg 2013).

Thirdly, from the results of Montana, SCC reports reveal disparities in TRI releases which are an important environmental hazard indicator for this study. Amount of releases of OSHA defined carcinogens chemicals in concerned Census Tracts is statistically much higher than the rest regions. Meanwhile, air quality in SCC areas is also tending to be lower than the rest regions. The distribution of impaired waters, TRI sites and concerned geographic areas seem fairly related. There might be some hidden chain reaction: TRI sites or other hazardous facilities contaminated the environment, and environmental issues might be later perceived by residents, which consequently triggered requests for cancer cluster investigations. Additionally, SCC reports also indicate difference of statistics for NURE sediment samples from non-concerned areas. Concentrations of both uranium and lead are higher in concerned geographic areas. Besides, according to the coefficient of logistic model, the concentration of lead plays the most importance role in triggering SCC reports. Trumbo and his colleagues argued that environmental hazards can bring impact on public perceptions of vulnerability to cancer. Although the disease incidence rate in certain communities are not very high, the fear of cancer is more prevalent in these communities located near contaminant sources, such as toxic exposure sites and hazardous waste sites (Trumbo 2008).

Lastly, the ArcGIS Online web application contributes to broaden the impact of this thesis (http://tinyurl.com/ppzkp6v). To my knowledge, there is no platforms provides visualized information about public SCC reports except for this web map. The interactive map allows users to gain knowledge and examine the geospatial relationship between SCC areas, potential hazards, authorative cancer data, SES status, etc. Besides, it sets up an extendable framework for more SCC data sharing in the future.

There are limitations as well as potential future works of this study. Firstly, the etiology of different types of cancer varies. If data is available, it would be more appropriate to divide SCC reports based on cancer types and explore the environmental factors separately Secondly, we encounter risks by using census tracts to integrate different SCC scales. When using multiple census tracts to represent a county or an American Indian Reservation area, there might be ecological fallacy. Because we assume every census tract has a SCC reports, which is not necessarily true. However, since SCC data do not show finer scale detailed information, the best we can do is deducting finer scale level data by assuming a homogeneous distribution. Thirdly, limitations to the TRI data are recognized: it does not provide human exposure information; it restricted to large manufacturing facilities (Fisher 2006); and it is self-reported dataset from facilities which might introduce noises and flaws to the dataset. Aside from environmental factors considered in this research, there are more environmental burdens such as Superfund sites, hazardous waste transfer, storage, and disposal facilities (TSDFs), access to parks and green spaces, noise pollution, etc. (Chakraborty 2011). Finer scale's authoritive cancer data and air quality data might be more preferable to against ecological fallacy. Fourthly, regarding to analytical methods, instead of constraining in predefined political boundary, pollutant fate and transport model can improve the accuracy of the assessment of the impact of hazards. The Risk-Screening Environmental Indicators (RSEI) and National Scale Air Toxic Assessment (NATA) are promising data sets derived from pollutant fate (Chakraborty 2011). Lastly, for web application, it would be ideally that users not only can explore the existing SCC reports but also participating in, for instance, making comment and feedback.

Admittedly, there are limitations in this study, but it does have contributions in public SCC studies. To conclude based on my samples, SCC reports do help to reveal disparities in terms of health, environmental contaminant. Results show that the demographic and socioeconomic characteristics are different in SCC areas.

REFERENCE

Abrams, B., H. Anderson, C. Blackmore, F. J. Bove, S. K. Condon, C. R. Eheman, J. Fagliano, L. B. Haynes, L. S. Lewis, J. Major, M. A. McGeehin, E. Simms, K. Sircar, J. Soler, M. Stanbury, S. M. Watkins & D. Wartenberg (2013) Investigating Suspected Cancer Clusters and Responding to Community Concerns Guidelines from CDC and the Council of State and Territorial Epidemiologists. *Mmwr Recommendations and Reports*, 62**,** 1-24.

ArcGIS. Create apps from maps.

ArcGIS Resources. How Spatial Autocorrelation (Global Moran's I) works.

ArcGIS Resources. 2012. ArcGIS Help 10.1 Euclidean Distance (Spatial Analyst).

Ayalew, L. & H. Yamagishi (2005) The application of GIS-based logistic regression for landslide susceptibility mapping in the Kakuda-Yahiko Mountains, Central Japan. *Geomorphology*, 65**,** 15-31.

Bogg, T. & B. W. Roberts (2013) The Case for Conscientiousness: Evidence and Implications for a Personality Trait Marker of Health and Longevity. Annals of Behavioral Medicine, 45, 278-288.

"Cancer_Cluster_Investigations." Cancer Cluster Investigations. Texas Department of State Health Services, 28 Jan. 2015. Web. 22 Apr. 2015. <http://www.dshs.state.tx.us/epitox/CancerClusters.shtm>.

CDC. U.S. Cancer Statistics: An Interactive Atlas.

CDC. 1990. Guidelines for Investigating Clusters of Health Events. 1-16. MMWR.

CDC. 2013. National Program of Cancer Registries (NPCR).

CDC. 2014. Black or African American Populations.

Chakraborty, J., J. A. Maantay & J. D. Brender (2011) Disproportionate Proximity to Environmental Health Hazards: Methods, Models, and Measurement. *American Journal of Public Health*, 101**,** S27-S36.

Community Health Book. 2014. THE MAP.

Corburn, J. (2003) Bringing local knowledge into environmental decision making - Improving urban planning for communities at risk. *Journal of Planning Education and Research*, 22**,** 420-433.

Corburn, J (2007) Community knowledge in environmental health science: co-producing policy expertise. *Environmental Science & Policy,* 10**,** 150-161.

Darden, J., M. Rahbar, L. Jezierski, M. Li & E. Velie (2010) The Measurement of Neighborhood Socioeconomic Characteristics and Black and White Residential Segregation in Metropolitan Detroit: Implications for the Study of Social Disparities in Health. *Annals of the Association of American Geographers,* 100**,** 137-158.

EPA. 2015. Lear about the Toxics Release Inventory. Toxics Release Inventory (TRI) Program.

Fan, J. X., M. Wen & L. Kowaleski-Jones (2014) An ecological analysis of environmental correlates of active commuting in urban U.S. *Health Place,* 30**,** 242-50.

Fisher, J. B., M. Kelly & J. Romm (2006) Scales of environmental justice: combining GIS and spatial analysis for air toxics in West Oakland, California. Health Place, 12, 701-14.

Florida Department of Health. 2009. Acreage Cancer Review Palm Beach County.

Friedman, H. S. (2000) Long-term relations of personality and health: Dynamisms, mechanisms, tropisms. Journal of Personality, 68, 1089-1107.

Author. 2013. State health official says Coldbrook Road cases in Hampden don’t indicate a cancer cluster. Nov 19.

Goodchild, M. (2007) Citizens as Sensors: the World of Volunteered Geography. *Geojournal,* 69**,** 211-221.

Goodchild, M. F., L. Anselin & U. Deichmann (1993) A Framework for the Areal Interpolation of Socioeconomic Data. *Environment and Planning A,* 25**,** 383-397.

Goodman, M., J. S. LaKind, J. A. Fagliano, T. L. Lash, J. L. Wiemels, D. M. Winn, C. Patel, J. Van Eenwyk, B. A. Kohler, E. F. Schisterman, P. Albert & D. R. Mattison (2014) Cancer cluster investigations: review of the past and proposals for the future. *Int J Environ Res Public Health,* 11**,** 1479-99.

Gordon-Larsen, P., M. C. Nelson, P. Page & B. M. Popkin (2006) Inequality in the built environment underlies key health disparities in physical activity and obesity. *Pediatrics,* 117**,** 417-424.

Greenberg, M. & D. Wartenberg (1991) Communicating to an alarmed community about cancer clusters: a fifty state survey. *J Community Health,* 16**,** 71-82.

Author. 2011. Group Calls for More Research into Camp Lejeune Cancer Cluster. *The Daily News* March 30.

IDRE. R Data Analysis Examples: Logit Regression. UCLA: Statistical Consulting Group.

James, G., D. Witten, T. Hastie & R. Tibshirani. 2013. *An introduction to statistical learning : with applications in R.*

Johnson, B. L. (2014) Notice of an Update of Guidelines on Investigations of Suspected Cancer Clusters. *Human and Ecological Risk Assessment*, 20**,** 287-289.

Jordan, C., C. S. Zhang & A. Higgins (2007) Using GIS and statistics to study influences of geology on probability features of surface soil geochemistry in Northern Ireland. *Journal of Geochemical Exploration*, 93**,** 135-152.

Juzych, N. S., B. Resnick, R. Streeter, J. Herbstman, J. Zablotsky, M. Fox & T. A. Burke (2007) Adequacy of state capacity to address noncommunicable disease clusters in the era of environmental public health tracking. *Am J Public Health*, 97 Suppl 1**,** S163-9.

Kingsley, B. S., K. L. Schmeichel & C. H. Rubin (2007a) An update on cancer cluster activities at the Centers for Disease Control and Prevention. *Environ Health Perspect*, 115**,** 165-71.

--- (2007b) An update on cancer cluster activities at the Centers for Disease Control and Prevention. *Environmental Health Perspectives*, 115**,** 165-171.

Lima, A., B. De Vivo, D. Cicchella, M. Cortini & S. Albanese (2003) Multifractal IDW interpolation and fractal filtering method in environmental studies: an application on regional stream sediments of (Italy), Campania region. *Applied Geochemistry*, 18**,** 1853-1865.

McComas, K. A., C. W. Trumbo & J. C. Besley (2007) Public meetings about suspected cancer clusters: The impact of voice, interactional justice, and risk perception on attendees' attitudes in six communities. *Journal of Health Communication*, 12**,** 527-549.

McGrath, D., C. S. Zhang & O. T. Carton (2004) Geostatistical analyses and hazard assessment on soil lead in Silvermines area, Ireland. *Environmental Pollution*, 127**,** 239-248.

Menard, S. W. 1995. *Applied logistic regression analysis*. Thousand Oaks, Calif.: Sage Publications.

Morello-Frosch, R., M. Pastor, C. Porras & J. Sadd (2002) Environmental justice and regional inequality in southern California: Implications for future research. *Environmental Health Perspectives*, 110**,** 149-154.

National Cancer Institue. 2014. Cancer Clusters.

Pastor, M., J. Sadd & J. Hipp (2001) Which came first? Toxic facilities, minority move-in, and environmental justice. Journal of Urban Affairs, 23, 1-21.

Pearce, J. R., E. A. Richardson, R. J. Mitchell & N. K. Shortt (2011) Environmental justice and health: a study of multiple environmental deprivation and geographical inequalities in health in New Zealand. Soc Sci Med, 73, 410-20.

Psychstatistics. 2012. Detecting Collinearity. Psychstatistics.

Reid, J. C. (1993) A Geochemical Atlas of North-Carolina, USA. *Journal of Geochemical Exploration,* 47**,** 11-27.

SEER. Overview of the SEER Program.

Author. 2001. Residents Near the Weldon Spring Cleanup Site Wonder What's Causing a String of Infant Deaths and Other Illnesses Rash of Cases Terrifies Neighbors, But Studies Usually Find No Links. Januray 7.

Sheppard, E., H. Leitner, R. B. McMaster & H. G. Tian (1999) GIS-based measures of environmental equity: Exploring their sensitivity and significance. *Journal of Exposure Analysis and Environmental Epidemiology,* 9**,** 18-28.

Shi, W. J., J. Y. Liu, Z. P. Du, Y. J. Song, C. F. Chen & T. X. Yue (2009) Surface modelling of soil pH. *Geoderma,* 150**,** 113-119.

Siegrist, M., G. T. Cvetkovich & H. Gutscher (2001) Shared values, social trust, and the perception of geographic cancer clusters. *Risk Analysis,* 21**,** 1047-1053.

Simpson, B. W., P. Truant & B. A. Resnick (2014) Stop and listen to the people: an enhanced approach to cancer cluster investigations. *Am J Public Health,* 104**,** 1204-8.

State Cancer Profiles. Quick Profiles for States.

Sui, D. Z., S. Elwood & M. F. Goodchild. 2013. *Crowdsourcing geographic knowledge : volunteered geographic information (VGI) in theory and practice*. Dordrecht ; New York: Springer.

Thun, M. J. & T. Sinks (2004) Understanding cancer clusters. *CA Cancer J Clin,* 54**,** 273-80.

Trumbo, C. W. (2000) Public requests for cancer cluster investigations: A survey of state health departments. American Journal of Public Health, 90, 1300-1302.

Trumbo, C. W., K. A. McComas & J. C. Besley (2008) Individual- and community-level effects on risk perception in cancer cluster investigations. Risk Analysis, 28, 161-178.

Wagner, S. E., J. B. Burch, M. Bottai, R. Puett, D. Porter, S. Bolick-Aldrich, T. Temples, R. C. Wilkerson, J. E. Vena & J. R. Hebert (2011) Groundwater uranium and cancer incidence in South Carolina. Cancer Causes & Control, 22, 41-50.

Whitman, S., A. Silva, A. Shah & D. Ansell (2004) Diversity and disparity: GIS and small-area analysis in six Chicago neighborhoods. *J Med Syst,* 28**,** 397-411.

Xie, Y. F., T. B. Chen, M. Lei, J. Yang, Q. J. Guo, B. Song & X. Y. Zhou (2011) Spatial distribution of soil heavy metal pollution estimated by different interpolation methods: Accuracy and uncertainty analysis. *Chemosphere,* 82**,** 468-476.

Yancik, R. (1997) Cancer burden in the aged - An epidemiologic and demographic overview. *Cancer,* 80**,** 1273-1283.

Yasrebi, J. S., Mahboub; Fathi, Hamed; Karimian, Najafali (2009) Evaluation and Comparision of Ordinary Kriging and Inverse Distance Weighting Methods for Prediction of Spatial Variability of Some Soil Chemical Parameters. *Research Journal of Biological Sciences,* 4**,** 93-102.

Zumlot, T., P. Goodell & F. Howari (2009) Geochemical mapping of New Mexico, USA, using stream sediment data. *Environmental Geology,* 58**,** 1479-1497.

APPENDIX List of abbreviations and acronyms

CDC – Centers for Disease Control and Prevention

EPA – Environmental Protection Agency

GIScience – Geographical Information Science

IARC – International Agency for Research on Cancer

IDW – Inverse distance weighting

NATA – National Scale Air Toxic Assessment

NCI – National Cancer Institute

NPCR – National Program of Cancer Registries

NURE – National Uranium Resources Evaluation Program

OLS – Ordinary least square regression

RSEI – The Risk-Screening Environmental Indicators

SEER – Surveillance, Epidemiology and End Results Program

SEP – Social economic position

SCC – Suspected cancer cluster

TRI – Toxics Releases Inventory

TSDF – Hazardous water transfer, storage, and disposal facilities

VIF – Variance inflation factor