Geographic Disparities in Cancer Mortality to Incidence Ratios

Kyle Buck
University of South Carolina - Columbia
Geographic Disparities in Cancer Mortality to Incidence Ratios

By

Kyle D. Buck

Bachelor of Science
University of South Florida, 2002

Master of Science
University of South Florida, 2011

Submitted in Partial Fulfillment of the Requirements

For the Degree of Doctor of Philosophy in

Geography

College of Arts and Sciences

University of South Carolina

2014

Accepted by:

Susan L. Cutter, Major Professor

Christopher T. Emrich, Committee Member

Diansheng Guo, Committee Member

Kellee White, Committee Member

Lacy Ford, Vice Provost and Dean of Graduate Studies
DEDICATION

To Kelly, Ethan, Collin and Emma. You have tolerated my obsession for the better part of six years. None of this would be possible without your love and support.
ABSTRACT

While cancer rates have shown promising trends over the last few decades, not all populations have experienced the same levels of decrease in cancer incidence in mortality rates. Identifying populations suffering from the impacts of the disparities has become a major goal in cancer research.

Most research has focused on the influence of single variables on cancer disparities or on small-scale case studies. Using the information from these analyses, the research conducted in this dissertation tests the relationship of selected variables to an outcome measure, the mortality to incidence ratio (MIR) in search of spatial relationships between the indicators and the MIR. The goal is to identify influential variables in addition to determining whether variables consistently express the same influence over the MIR.

In order to achieve the goal, three separate analyses are conducted to correspond with the primary research questions. To answer the first research question, involving the identification of predominant socio-spatial indicators driving cancer disparities in the US, a regression model is run, using thirty-four potential variables as independent variables and the MIR as a dependent variable. The second research question, based on the identification of broad-based factors accounting for disparities in cancer outcomes, is answered through both theoretical and inductive grouping of the indicators using a priori knowledge and principle components analysis. A second regression tests the predictive
ability of the two grouping methods and the contribution of each group to the MIR. A path analysis is conducted as well to determine how factors influence each other and interact to yield cancer outcomes. The third research question, intending to identify whether the relationship of the broad-based factors to cancer outcomes remains consistent across the United States, is conducted using spatial methodologies. This final step involves a combination of hot spot mapping, geographically weighted regression analysis, and a bivariate Moran’s I to establish regions where disparities exist as well as identifying differences in the contribution of variables to the disparities.

The findings of the research reveal a complex interaction of variables and a level of dependence between the aggregated groups. The results of the first research question revealed obesity as the most highly correlated indicator to the MIR. Counties with higher rural populations were second, while social indicators including the percentage of single parent households and unmarried population also factored very highly into the model. When the indicators were grouped via theoretical models, health and behavioral characteristics along with social characteristics displayed most of the variance and had the highest correspondence to the MIR. The final research question, looking for spatial patterns, revealed a significant hot spot in the Southeast United States for both the MIR, social, health, and access factors. Similarities are most evident between the spatial patterns of the MIR in comparison to social and health characteristics. With the presence of definitive regional patterns and clear connections between the MIR and societal groupings, the finding from this research suggest a need to shift to sub-regional analysis in order to determine whether the same patterns hold up at a local level.
# TABLE OF CONTENTS

DEDICATION ........................................................................................................ iii

ABSTRACT ........................................................................................................... iv

LIST OF TABLES ................................................................................................... vii

LIST OF FIGURES ............................................................................................... ix

CHAPTER 1: INTRODUCTION .............................................................................. 1
  1.1 Overview ........................................................................................................ 1
  1.2 Research Objectives ..................................................................................... 3
  1.3 Structure of the Document .......................................................................... 6

CHAPTER 2: CONCEPTUAL DEVELOPMENT .................................................. 7
  2.1 Defining Cancer Vulnerability ..................................................................... 7
  2.2 Conceptual Development of Vulnerability in Hazards Geography ............. 8
  2.3 Conceptual Development in Health Disparities ......................................... 12
  2.4 Place-based Health Vulnerability Model .................................................... 15

CHAPTER 3: MEASUREMENT OF DRIVERS .................................................. 18
  3.1. Data Sources ............................................................................................ 20
  3.2. Initial Vulnerability Drivers ...................................................................... 20
  3.3. Cancer Outcome Indicator - Mortality to Incidence Ratio (MIR) ............... 43
  3.4. Summary and Model Operationalization .................................................... 45

CHAPTER 4: METHODOLOGY ............................................................................ 48
4.1. Study Region........................................................................................................... 48
4.2. Pre-processing of Data............................................................................................ 49
4.3. Procedures.............................................................................................................. 50

CHAPTER 5: PREDOMINANT VULNERABILITY FACTORS ......................... 61
5.1. Predictive Ability and Model Adequacy................................................................. 61
5.2. Identifying Significant Indicators ......................................................................... 64

CHAPTER 6: IDENTIFICATION OF FACTORS .............................................. 69
6.1. Factor Creation through Theoretical Grouping of Variables ............................... 69
6.2 Factor Creation through PCA Grouping of Variables ............................................ 75
6.3. Summary of Group Findings .................................................................................. 84
6.4. Verification of Group Influence on MIR – Path Analysis .................................... 92

CHAPTER 7: SPATIAL VARIABILITY OF VULNERABILITY FACTORS ....... 100
7.1. Hot Spot Mapping ............................................................................................... 100
7.2. Bivariate Moran's I ............................................................................................. 125
7.3. Geographically Weighted Regression ................................................................... 130
7.4. Summary of Visual Analysis ............................................................................... 134

CHAPTER 8: CONCLUSIONS ........................................................................... 136
8.1. Research Question 1 ............................................................................................ 137
8.2. Research Question 2 ............................................................................................ 139
8.3. Research Question 3 ............................................................................................ 140
8.4. Research Summary and Significance ................................................................... 142
8.5. Research Applications ......................................................................................... 143
8.6. Caveats ......................................................................................................................... 147

REFERENCES ..................................................................................................................... 150

APPENDIX A. DETAILED DATA LIST ........................................................................ 159
LIST OF TABLES

Table 3.1. Detailed list of initial vulnerability indicators and sources.........................21
Table 5.1. Regression results for 34-variable model.................................................61
Table 5.2. Regression model statistics for 34 variables.............................................63
Table 5.3. Significant MIR correlations in 34-variable regression model.......................65
Table 6.1a. Pearson's R for 34 grouped variables....................................................70
Table 6.1b. Pearson's R for 4 theoretically grouped factors.....................................71
Table 6.2. Theoretically Grouped Factors Regression model results............................72
Table 6.3. Theoretically Grouped Factors Explanatory variables...............................72
Table 6.4. Cronbach's Alpha for theoretically grouped factors..................................73
Table 6.5. Results of PCA run on 34 variables..........................................................77
Table 6.6. Paired t-test for access characteristics factor and PCA factor 1...................78
Table 6.7. Paired t-test for health and social factors with PCA factor 2......................79
Table 6.8. Paired t-test for social, access and health factors with PCA factor 3..........80
Table 6.9. Regression results for PCA-grouped variables..........................................82
Table 6.10. Path analysis scores calculated...............................................................96
Table 7.1. Results of GWR analysis on theoretically grouped factors.........................126
Table 7.2. Results of GWR analysis on PCA grouped factors....................................130
Table 7.3. Results of GWR analysis on theoretically grouped factors.......................131
Table 7.4. Results of GWR analysis on PCA grouped factors....................................134
Table A.1: Detailed variable list with computations and source URL.......................158
LIST OF FIGURES

Figure 2.1. Hazards-of-Place Model of Vulnerability (Cutter 1996).................................11
Figure 2.2. Health Disparities Conceptual Models.................................................................14
Figure 2.3. Place-based health vulnerability model.................................................................16
Figure 4.1. Flow diagram for aspatial validation procedure......................................................51
Figure 6.1. The 4 theoretical groupings along with their contribution to the MIR...............86
Figure 6.2. Diagram depicting the 10 PCA constructed factors.............................................87
Figure 6.3. Path model for theoretically grouped factors with the MIR.................................93
Figure 6.4. Path diagram with community factor removed.....................................................97
Figure 6.5. Path diagram with correlation and community factor removed.........................98
Figure 7.1. MIR Hot Spot Analysis. Getis Ords G* statistic..................................................101
Figure 7.2. Social Characteristics Hot Spots. Getis Ords Gi* for social disparities.........103
Figure 7.3. Single Parent Households and Non-white population hot spot maps.............104
Figure 7.4. Health and Behavioral Hot Spots. Getis Ords G*...............................................105
Figure 7.5. Economic Hot Spots. Getis Ords G* for financial disparities.........................107
Figure 7.6. Education Hot Spots. Getis Ords G*.................................................................108
Figure 7.7 - Community Hot Spots. Getis Ords G*............................................................110
Figure 7.8. PCA Factor 1 Hot Spots. Getis Ords G* for Factor 1 values.........................111
Figure 7.9. PCA Factor 2 Hot Spots. Getis Ords G* for Factor 2 values.........................113
Figure 7.10. PCA Factor 3 Hot Spots. Getis Ords G* for Factor 3 values.......................114
Figure 7.11. PCA Factor 4 Hot Spots. Getis Ords G* for Factor 4 values.......................116
Figure 7.12. PCA Factor 5 Hot Spots. Getis Ords G* for Factor 5 values.................117
Figure 7.13. PCA Factor 6 Hot Spots. Getis Ords G* for Factor 6 values.................119
Figure 7.14. PCA Factor 7 Hot Spots. Getis Ords G* for Factor 7 values...............120
Figure 7.15. PCA Factor 8 Hot Spots. Getis Ords G* for Factor 8 values...............121
Figure 7.16. PCA Factor 9 Hot Spots. Getis Ords G* for Factor 9 values...............122
Figure 7.17. PCA Factor 10 Hot Spots. Getis Ords G* for Factor 10 values.............123
Figure 7.18. Bivariate Moran’s I maps. MIR with each of the theoretical groups……125
Figure 7.19. Bivariate Moran’s I maps comparing MIR with inductive groups 1-5…..128
Figure 7.20. Bivariate Moran’s I maps comparing MIR with inductive groups 6-10…129
Figure 7.21. GWR residuals for theoretically grouped factors.................................132
Figure 7.22. GWR residuals for PCA grouped factors............................................134
Figure 8.1. Revised concept map based on results presented.................................142
Figure 8.2. Conceptual Model created using U.S. county-level data.......................144
CHAPTER 1

INTRODUCTION

1.1. Overview

The impact of cancer is enormous and takes a toll on both the individual and societal level. The economic impact is estimated at $228 billion dollars between both direct health costs and indirect costs from premature death (American Cancer Society 2012). This burden comes in the form of nearly 12 million people living with cancer in the US as well as an annual mortality of over 500,000 (Howlander, Noone, and Krapcho 2010). There is good news amidst the bad, however. Cancer incidence and mortality rates have been dropping in recent years according to the American Cancer Society. In their 2010 release of Cancer Facts and Figures, the American Cancer Society estimated that death rates fell by 21% for men and 12.3% for women between 1991 and 2006. The incidence rates also fell, although not by nearly as impressive a margin, 1.3% for men and 0.5% for women in the same period (American Cancer Society 2010).

At the same time as the release of the American Cancer Society facts and figures, the National Institute for Health (NIH) released a very different report on health disparities as a part of its Healthy People 2010 campaign. The exposed gap in cancer outcomes among social groups was wide enough to merit several in-depth investigations by National Cancer Institute (NCI) researchers, and resulted in a call for further attention to drivers of these disparities. Although the overall picture of cancer impact in the US looks brighter, the effect is not felt equally among all groups in the US. Cancer health
disparities, as defined by NCI, are “adverse differences in cancer incidence, cancer prevalence, cancer death, cancer survivorship, and burden of cancer or related health conditions that exist among specific population groups in the United States” (United States Public Law 106-525 (2000), p. 2498).

A major goal in the realm of cancer research involves the elimination of these cancer-related health disparities, which result in diverse rates of incidence and mortality, as well as differences in the quality of life. Factors instigating a divide among the groups range from health factors to environmental exposures to social processes resulting in differential diagnoses and treatments for the disease. The NCI has funded numerous programs and research initiatives aimed at the measurement and remedy of these existing inequalities. Their concern lies in the lack of cohesive analysis, stating that, “despite the increased attention to social disparities in health, no clear framework exists to define and measure health disparities” (Harper and Lynch 2010). Filling in this gap, creating a framework to measure health disparities is a major component of this research. By shifting the conceptual model for the measurement of health disparities, this research contributes through offering a concrete method to assess geographic disparities and identify populations at a national scale.

Research thus far has typically focused on the measurement and comparison of rates, such as incidence and mortality (outcome measures), to one of many socioeconomic variables. The problem with this approach is the one-dimensional manner in which the social predictors are handled. Each variable is compared to the cancer outcomes as if it exists independently from other variables. For instance, testing income levels against incidence rates for prostate cancer could identify certain patterns. It would
not account for the interaction between income and other elements of potential influence over the same outcome. In order to explore the interaction between drivers and their cumulative impact on cancer outcomes, a different method must be employed that allows for all factors to be tested simultaneously against the measure of interest. Analysis of geographic disparities along with the use of a comprehensive cancer indicator allows for streamlined identification of disparity populations as well as identification of potential drivers of the disparities.

1.2. Research Objectives

The goal of this research is to identify the primary factors influencing cancer-related health disparities, the relative impact of these factors on the disparities, and the spatial extent of these factors on a county level. The motivation for this research stems from a lack of reliable, place-based indicators that can be used to identify populations with higher vulnerability to negative cancer outcomes. A robust set of cancer vulnerability drivers, corresponding to known outcome measures, creates a more complete picture of the populations that are not seeing the same improvements in cancer outcomes. Knowing the factors responsible for driving the disparities in those areas ultimately translates to more effective prevention and treatment, thus helping to close the gap. Throughout the dissertation, indicators refer to specific variables and factors refer to the aggregation of the indicators into specific thematic areas. The following research questions provide the impetus for this dissertation.
1. What are the predominant socio-spatial indicators driving cancer disparities in the US?

2. Which of the broad-based factors account for most of the spatial variability in cancer outcomes, (social, financial and medical access, community and environmental, or health and behavioral)?

3. Will the relationship of the broad-based factors to cancer outcomes remain consistent across the United States?

There are currently a number of gaps in the research and this study seeks to fill those gaps. The first gap, as stated previously, is the lack of a consistent framework to identify and measure health disparities. The contribution provided by this research is related to the spatial analysis of cancer disparities. Identification of geographic disparities offers a simpler and more concrete method in addition to creating a visual representation of patterns.

The second gap lies in the use of either incidence or mortality rates as the outcome measure. Cancer vulnerability involves more than simple incidence, mortality, or prevalence alone. While each of these measures is helpful in identifying aspects of population health, there are many theoretical issues involved in using any as a measure of cancer disparities. Cancer latency periods can be very long and people tend to relocate, making it difficult to ascertain what factors initiated the cancer. Residents also may have been exposed to the factors prior to their relocation. Mortality rates offer a more reliable approach to the measurement of cancer vulnerability, but not without their issues. Relocation will have the same impact on mortality rates as it does on incidence. Mortality is typically the preferred outcome measure because it tends to reflect stage of diagnosis.
and treatment better than incidence. Dividing the mortality by the incidence rate creates mortality to incidence ratios (MIRs) for each of the study areas, and eliminates the problems of latency period and other mobility issues related to human populations. This measure will also control for the variance in the incidence rates and provide a more robust indicator of cancer disparities. Documentation of the geographic patterns of the MIR is an important first step in identifying regions for future research. Also an important consideration is the type of burden that is present in an area.

Incidence rates are useful to highlight areas with higher cancer burdens, indicating areas that may require more medical personnel or special treatment centers to handle the number of cases. There also may be underlying social or environmental factors that are connected to the higher incidence rates that do not show up when analyzing the MIR. Both scenarios would prove valuable to the entities responsible for allocation of resources and to the potential alleviate of the disparities. Mortality rates could also shed light on a few potential issues at the sub-regional level. Higher incidence of poor-prognosis cancer could lead to higher mortality. Mortality may also expose areas with cultural barriers, monetary challenges, or issues with access that lead to inadequate treatment. These are all preventable causes, but must be identified in order to take action.

Using the MIR as a method to identify disparity populations at a national level is a significant contribution of this research. The MIR is used as an outcome measure due to its accuracy in capturing both the early detection of cancer and any effective treatment outcomes. Also, due to the interest in cancer disparities, the MIR is used to help isolate counties that are not receiving appropriate care, most likely due to differences in SES (Hebert 2009). Using this ratio also removes the question of latency period and
geographic relocation that would be present when using either incidence or mortality rates alone.

1.3. Structure of the Document

This dissertation is composed of six further chapters, each intended to address a distinct component of measurement and identification of cancer disparities. In chapter 2, the conceptual underpinnings of health vulnerability are analyzed from the perspective of hazards geography. The purpose of this chapter is to establish the theory and model that the rest of the research will be based upon. Chapter 3 answers the first research question, to identify individual characteristics that have more influence on the mortality to incidence ratio. Chapter 4 addresses the broad-based factors influencing cancer disparities. The concern of this piece is to identify the variables with the most variability as well as to determine how the variables relate to one another. This helps to answer the question of how variables influence each other, both negatively and positively. Chapter 5 involved the spatial analysis, looking for trends in the data from place to place. This also analyzes the relationship of the identified factors to the outcome measure to flesh out the pattern of interactions. The final chapter serves as a synthesis of the previous three chapters and clearly ties the concepts together while also commenting on the operationalization of the conceptual model.
CHAPTER 2

CONCEPTUAL DEVELOPMENT

There are two major fields of research merged in this study. Health disparities research drives a portion of the theoretical factor grouping. Each of the four main disparities models guides the creation of the four a priori groups. Hazards geography research drives both a portion of the theoretical factor grouping and the development of a spatial model to measure the cancer disparities. The field of health disparities has developed in a similar manner to that of hazards geography. While certain biological and physical factors influence health outcomes, social factors also play a very large role in determining which groups will be exposed and the level of treatment they will receive. This critical link between the hazards and health research provides the foundation for this research.

2.1. Defining Cancer Vulnerability

Establishing a clear and consistent definition of vulnerability is an essential first step in this research. Broadly stated, within the field of hazards geography, vulnerability is the potential for loss. Within health disparities research, vulnerability defined as the susceptibility to harm. Harm, in this case, would be negative health outcomes. While the definitions appear simple on the surface, pinning down the causes results in substantial complications. Hazards geography defines the causes as natural or man-made events.
Natural events are relatively simple, but the man-made causes begin to encompass the complexities of the social environment. In health disparities, the agreed upon factors are a combination of exposure and social variables, much like hazards geography. In a study of vulnerability in health care, Rogers defines a third factor influencing vulnerability (1997). She refers to vulnerability as situational, meaning a person’s vulnerability can change depending on their environment. This conceptualization of vulnerability shares a strong similarity to that used within the hazards of place-based model in geography and lends well to the merging of the two definitions for use in a new model. People have inherent components making them vulnerable, but these components vary spatially and temporally. For the purposes of this research, cancer vulnerability is defined as the potential for loss among a group in a specific geographic area resulting from a combination of health, socio-behavioral, and environmental factors.

Identification of vulnerable populations ties together hazards research and health disparities research. Each field has developed models and metrics for identification of these vulnerable populations over the years, ending up with products that share important similarities. In order to effectively combine the ideas of the two, it is important to first look at the development of the concepts within each field.

2.2. Conceptual Development of Vulnerability in Hazards Geography

Current concepts and metrics from the field of hazards geography can make relevant contributions to the field of health disparities by integrating the complex environmental and social systems that have thus far complicated the measurement of vulnerability. In order to effectively measure such complex systems, the disparities research must first be framed using a more suitable conceptual framework--one that takes
into account the interaction of social networks with behavioral and environmental factors influencing health outcomes. The hazards of place model of vulnerability proposed provides an excellent starting point in this regard by accounting for the interactions of environmental and social systems as they vary in space and time (Cutter 1996). By focusing on the place as a fundamental unit of analysis, this model allows for an investigation into the interaction of driving factors and their impact on health outcomes.

Looking first at the conceptual models in geography, vulnerability is closely tied with risk and disasters. In order for a population to be vulnerable, there must be a hazard, or risk, present. This is the inherent risk presented by the environment, where measurement of exposure to a hazard serves as the proxy. Vulnerability is more refined, considering both the exposure to a hazard and the modification of this risk by social factors. Combined, these two yield a net vulnerability for a population. The term vulnerability has been in use by geographers since the 1980s, with Peter Timmerman offering one of the first conceptualizations in 1981 (Timmerman 1981). Although complete consensus on the meaning of the term still has not been reached, there have emerged three main tenets of vulnerability in geographic research. Vulnerability is defined as a potential for loss resulting from a combination of exposure (Burton, Kates, and White 1993; Anderson 2000), an underlying social condition (Wisner 2004), and the combination of the two as they play out in different geographic areas (Kasperson, Kasperson, and Turner 1995; Cutter, Mitchell, and Scott 2000). Some recent vulnerability research has espoused a synthesis of the three lineages (Eakin and Leurs 2006). In this combination, the complexities of society are more accurately addressed and new insights are more likely.
Taking vulnerability research a step further, Cutter (1996) published the hazards of place model of vulnerability, built on the concept of the hazardousness of place (Hewitt and Burton 1971). In this model, vulnerability is measured as a combination of exposures and social influences on a specific place and time. The vulnerability can change depending on exposure to hazards and mitigation efforts, and is represented as the vulnerability of a place. The benefit of a model such as this is the ability to scale it depending on needs. The place vulnerability, as shown in Figure 2.1, could represent anywhere from an entire country down to a census block tract.

The social vulnerability piece of the model takes into account the complexities of the societal construct, recognizing its contribution to overall vulnerability. Even when the biophysical exposure is identical, different individuals or populations will experience different levels of vulnerability based on a number of social factors present. In addition, the social fabric could be a mitigating force, in the cases where planning or other modifications are enacted to reduce the risk of loss.

In geography, biophysical vulnerability is typically measured as the distribution and exposure of hazards and the degree of loss from hazard events. Social vulnerability is more complex; dealing with the intersection of multiple stressors and social forces in a place. Put together, the vulnerability of a place is a composite of both the biophysical and the social vulnerability present.
The Hazards of Place Model was downscaled and operationalized in Georgetown County, South Carolina (Cutter et al. 2000). While this case study provided valuable insight into place vulnerability, the social piece of the model did not pick up much of the local variability. This lacking led to the development of a place-based and inductive approach to measuring the social vulnerability of a place.

The empirically based method developed for the identification and analysis of social variables influencing vulnerability to hazards is called the Social Vulnerability Index (SoVI) (Cutter, Boruff, and Shirley 2003). Since this time, the Social Vulnerability Index (SoVI) has been used in many instances as a method for empirically based identification of groups vulnerable to greater losses in the event of a hazard.

Applications have included assessments of vulnerability to natural disasters as well as recent discussions that link vulnerability with the concepts of recovery and resilience (Cutter et al. 2008; Cutter, Burton and Emrich 2010). The SoVI model is able to capture
the factors contributing to social vulnerability in each county and proves that these factors vary through space and time. This model can be replicated for use in health disparities research in an attempt to capture the complexities of the social infrastructure leading to the disparities.

Vulnerability to negative health outcomes such as cancer, much like vulnerability to hazards, can be defined as the potential for loss. The loss could be measured as death or the social and financial burdens of cancer treatment. The factors influencing loss would also be the same as in hazards research, considering a combination of exposure, underlying social condition, and the manner in which these two play out in a geographic area. Exposures are comprised of both social and behavioral factors known to increase the risk of cancer incidence and mortality, including smoking, hazardous occupations, and pollution. Social considerations include education level, income, and access to care.

2.3. Conceptual Development in Health Disparities

Research into health disparities has garnered more attention recently with regards to both the measurement of disparities and the determination of their drivers. Many recent analyses have focused on the impact of socioeconomic status and race on disparities in asthma (Gold and Wright 2005), diabetes (Peek, Cargill and Huang 2007), cardiovascular disease (Davis et al. 2007), and infant mortality in addition to cancer (MacDorman et al. 2007). Others have looked into the relative impact social characteristics on general health disparities (Do, Frank and Finch 2012). The Prevention Institute also made an attempt at measuring disparities in communities and identifying areas of vulnerability and resilience to negative health outcomes. Their research, referred to as the Toolkit for Health & Resilience In Vulnerable Environments (THRIIVE),
provides a broad overview of the numerous influences within a community and how these affect health (Prevention Institute 2004).

Health disparities research has thus far failed to come to an agreement regarding the appropriate methodology used to identify disparity populations. This has limited the ability to effectively move forward and begin work on the representation of complex relationships among variables and variance in disparities from place to place. Using the geographic model proposed here allows for a more complete evaluation of disparities due to the use of place as the unit of analysis. In merging the conceptual models of health disparities and hazards geography, all factors can be assessed simultaneously within a geographic area.

Much of the methodological variation found in health disparity research stems from a lack of agreement over the appropriate conceptual model to follow. Currently, there are four predominant models existing in disparities research, each based on a different set of sources believed to drive health outcomes (Roux 2012). Current trends in the field call for a merging of the models to form an integrated, complex systems approach to modeling health disparities. Health inequalities are considered preventable, and a consolidation of models would allow for easier application of current knowledge in future research. Models reflecting the more realistic, complex interaction evident in society will ultimately yield more accurate assessments of the disparities and drivers causing them.

Conceptual models within the health disparities realm vary based on the predictive factors seen as most influential. In 2012, Ana V. Diez Roux conducted
research into four primary examples within the health disparities field: the genetic, fundamental-cause, pathways, and interaction models, shown in Figure 2.2 (Roux 2012). Each of these models is responsible for different aspects of the research conducted within health disparities. The genetic model is based on perceived genetic differences in population and serves as the construct for studies centered on racial or ethnic differences. The fundamental cause model looks at structural factors related to social and economic organization. This model drives studies related to socioeconomic status (SES) and health disparities. The pathways model highlights mediating pathways, or links between factors and outcomes. The initial factors of vulnerability are still important, but the focus is on the mechanisms that cause disparities as a result of the initial factors. For example, this model would drive an investigation into events that would stem from having a lower education level, assuming that education level sets in motion a chain of events leading to higher vulnerability.

![Health Disparities Conceptual Models](image)

Figure 2.2. Health Disparities Conceptual Models. (a) Genetic Model (b) Fundamental-Cause Model (c) Pathways Model (d) Interaction Model.
The interaction model focuses on the presence of multiple factors that interact to lead to disparities, notably the gene-environment interaction. Some factors, when combined, will cause a higher degree of vulnerability, while others will serve to decrease the vulnerability. Diez Roux maintains that the four models are interrelated and complementary and each serves to explain a piece of the larger puzzle leading to disparities in health outcomes. The implied challenge is to create a new analytic approach and tool set allowing the four models to be combined, thus creating a more robust approach to measuring the complex relationships that spur health disparities.

2.4. Place-based Health Vulnerability Model

The proposed conceptual model of place-based health vulnerability forms the backbone of this research and is significant in its combination of the spatial methodologies adopted from hazards geography with the health disparities models. Figure 2.3 demonstrates a synthesis of the four health disparities models within the framework of the Hazards of Place model. By breaking apart each of the components of health risk, it is now possible to operationalization and measure the influence of each component. The resulting health vulnerability is a sum of the components within a given geographic space. As in the Hazards of Place model, there is the potential for scaling to accommodate the analytical needs.

In this conceptualization, vulnerability begins with the underlying health and behavioral characteristics as well as financial access to medical care, which interact to yield a baseline health risk. In this model, underlying risk is a gauge of the negative health outcomes arising from existing health conditions, measured by obesity rates, poor health ratings and low birth weight children born into the population, to name a few.
Also adding to the health risk component are the access characteristics of a place. These variables measure proximity to medical care based on both the per capita physical availability of facilities and doctors as well as the financial ability of the population to afford the care. They include indicators such as the number of screening facilities, number of doctors (general practice and internists), unemployment rate, median household income, and the percentage of the population with health insurance. The resulting health risk is filtered through the interaction of social fabric to yield social vulnerability. In a similar fashion, the health risk is filtered through community characteristics to yield environmental vulnerability. These include access to parks and

Figure 2.3. Place-based health vulnerability model.
recreational facilities, population density, access to liquor and healthy food stores, fast food restaurants, and high-risk occupations. There is an interaction shown between the social fabric and the community characteristics indicating a close relationship between the two constructs. Each factor will influence the other, and contribute to changes in the health vulnerability of a place. In this model, the shift in terminology from risk to vulnerability marks the change to a place-based measurement, rather than an individual-based measure.

A big piece of this research lies in the correspondence of the health disparities and hazards geography fields and what they are attempting to measure. Establishing the connection based on the concept of vulnerability provides a justification for the combination of fields as well as the formation of a conceptual model merging the two. The ability to operationalize the model is of key concern in this research, as it allows for the identification and measurement of cancer disparities based on place and the measurement and comparison of the constructed factors to the places with identified disparities.
CHAPTER 3 

MEASUREMENT OF VULNERABILITY DRIVERS

Within the field of hazards geography, a great deal of research has been conducted on drivers of social vulnerability, with great attention paid to this interaction (Adger 2006; Cutter, Mitchell and Scott 2000; Cutter, Boruff and Shirley 2003). What the hazards research has revealed is an intricate social structure with a high geographic dependence. Another finding is that the spatial variance of many societal indicators is significant and most likely influences the vulnerability of populations to loss. One social factor does not always exert the same level of influence on vulnerability. Utilizing the knowledge gained in the hazards field provides a much better metric for assessment of vulnerability to negative cancer outcomes. The outcomes as well as the drivers of vulnerability between cancer and hazards are very similar and treating the analysis of them similarly is a logical progression in the advancement of cancer outcomes prediction.

Social factors influencing cancer-related health outcomes are well documented. Variables such as gender, age, ethnicity, education, income, disability status, health care accessibility, and occupation are frequently cited as drivers for a multitude of health risks. Combinations of factors have been utilized in a few studies, but the scale has remained limited and only a small number of variables are used in each case (Wagner et al. 2012; Li, Sundquist and Sundquist 2012).
Sam Harper and John Lynch produced a research paper on the measurement of cancer health disparities in 2010 (Harper and Lynch 2010). In it, the researchers focus on how to best quantify disparities in cancer outcomes. A multitude of social factors are utilized in this effort to define groups and identify the levels of disparity existing in the landscape. Their research is an attempt to identify the best measures of disparity in order to more effectively track and ensure their elimination.

Health disparities can stem from ethnic, gender, income, and age divisions. In order to accurately reflect the influence of these, the analysis must account for multiple combinations of variables that can exist amongst groups. It is not necessarily accurate to say a group is of a certain social class, and therefore more vulnerable. Other social indicators may exist, making them more or less vulnerable. For example, an individual may be vulnerable due to their age, but this vulnerability could be decreased if the individual is a wealthy, married female. Determining the relative impact of all cancer drivers in addition to how these drivers interact with each other will allow for a much more thorough and accurate assessment of the social landscape and lead to better measurement of the drivers.

Cancer is chosen as a health outcome for a multitude of reasons. As established earlier, cancer places an enormous burden on people in the United States. Also, cancer has a well-researched history and established patterns of disparities among certain populations. Finally, cancer data is of high quality and is publicly available countrywide, with most states maintaining cancer registries.
3.1. Data Sources

All data collected for this research is freely available and accessible on the national scale. Most indicators are collected from the Decennial Census, Economic Census, and the American Community Survey. Behavioral indicators are collected from the Center for Disease Control’s Behavioral Risk Factor Surveillance System (BRFSS) and the University of Wisconsin’s County Health Rankings. The BRFSS conducts interviews and questionnaires about health factors, and compiles the results at the county level. Information regarding number of doctors and facilities comes from the Area Resource File. The temporal availability of each indicator lies in the range of 2005-2010. Every attempt has been made to match the date of data collection within each of the indicator sets.

Data for outcome measures is obtained from the Center for Disease Control’s National Program of Cancer Registries (NPCR). NPCR data is collected both for incidence and mortality rates and used to calculate a Mortality-to-Incidence Ratio (MIR) (Hebert et al. 2009). For more detail on the MIR and its construction see section 3.3.

3.2. Initial Vulnerability Drivers

The literature identifies a number of indicators responsible for adverse health impacts, shown in table 3.1. In this study, the collected indicators are grouped into four sets, or factors, shown to impact a population’s health vulnerability. They are social characteristics, general health and behavioral characteristics, financial and medical access characteristics, and community and environmental characteristics.

This section will describe the thirty-six individual indicators used to measure the population vulnerability. There is an a priori assumption that each indicator will
contribute to the vulnerability of in a certain way. This assumption will be tested in order to answer the first and second research questions posed. A more detailed table is also available in Appendix A. This table provides the location of all data sources as well as the computations used, if any, to derive the values.

Table 3.1. Detailed list of initial vulnerability indicators and sources.

<table>
<thead>
<tr>
<th>Predictor Variables</th>
<th>Data Source</th>
<th>Year Available</th>
<th>Calculations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outcome measure: MIR</td>
<td>CDC-NPCR</td>
<td>05-09</td>
<td>Calculated as the mortality rate divided by the incidence rate</td>
</tr>
<tr>
<td>Income (-)</td>
<td>Census - ACS</td>
<td>05-09</td>
<td>Mean household income in last 12 months</td>
</tr>
<tr>
<td>Income Inequality (+)</td>
<td>County Health Rankings</td>
<td>05-09</td>
<td>GINI Index – Income inequality Range (0-1)</td>
</tr>
<tr>
<td>Unemployed (+)</td>
<td>Census - ACS</td>
<td>05-09</td>
<td>Percentage unemployed</td>
</tr>
<tr>
<td>Renters (+)</td>
<td>Census - ACS</td>
<td>05-09</td>
<td>Tenure- calc. percentage of renters</td>
</tr>
<tr>
<td>Race-Non-white (+)</td>
<td>Census - ACS</td>
<td>05-09</td>
<td>Percentage of population not classified as white</td>
</tr>
<tr>
<td>Religious affiliation (-)</td>
<td>US Religious Census</td>
<td>2010</td>
<td>County level congregation membership</td>
</tr>
<tr>
<td>Married population (-)</td>
<td>Census - ACS</td>
<td>05-09</td>
<td>Percentage of pop (&gt;18) now married</td>
</tr>
<tr>
<td>Single-parent household (+)</td>
<td>Census - ACS</td>
<td>05-09</td>
<td>Male householder + female householder</td>
</tr>
<tr>
<td>Number of dependents (+)</td>
<td>Census - ACS</td>
<td>05-09</td>
<td>Percentage of families with &gt;1 dependent (&lt;18 or &gt;65 years old)</td>
</tr>
<tr>
<td>Educational Attainment (-)</td>
<td>County Health Rankings</td>
<td>2010</td>
<td>Percentage of population have a college diploma</td>
</tr>
<tr>
<td>Language isolation (+)</td>
<td>Census - ACS</td>
<td>05-09</td>
<td>Percentage of population not speaking English</td>
</tr>
<tr>
<td>Parks per thousand (-)</td>
<td>USDA-ERS</td>
<td>2010</td>
<td>Count of all parks standardized by county population</td>
</tr>
<tr>
<td>Recreation Facilities (-)</td>
<td>County Health Rankings</td>
<td>2010</td>
<td>Per capita count of recreational facilities in a county.</td>
</tr>
<tr>
<td>Natural Amenities Scale (-)</td>
<td>USDA-ERS</td>
<td>1999</td>
<td>Index for livability of area based on climate factors</td>
</tr>
<tr>
<td>Environmental hazards (+)</td>
<td>EPA-TRI Locator</td>
<td>2009</td>
<td>Total amount of emissions from TRIs in county per capita</td>
</tr>
<tr>
<td>Rural population (+)</td>
<td>Census 2010</td>
<td>2010</td>
<td>% Living in rural areas - calculated</td>
</tr>
<tr>
<td>Particulate Matter Days (+)</td>
<td>EPA- County Health Rankings</td>
<td>2010</td>
<td>Number of days the particulate latter exceeded safe limits</td>
</tr>
<tr>
<td>Ozone Days (+)</td>
<td>EPA- County Health Rankings</td>
<td>2010</td>
<td>Number of days the level of ozone exceeded safe levels</td>
</tr>
<tr>
<td>Liquor Store Density (+)</td>
<td>County Health Rankings</td>
<td>2010</td>
<td>Density of liquor stores per square mile in the county</td>
</tr>
<tr>
<td>Fast Food Access (+)</td>
<td>USDA-ERS</td>
<td>2010</td>
<td>Number of fast food restaurants per 1000 population</td>
</tr>
<tr>
<td>High risk occupation (+)</td>
<td>Economic Census</td>
<td>2009</td>
<td>Percentage working in high risk occupations</td>
</tr>
<tr>
<td>Health Food Access (-)</td>
<td>County Health Rankings</td>
<td>2010</td>
<td>Percentage of zip codes in county with healthy food options</td>
</tr>
<tr>
<td>Population density (+)</td>
<td>Census - ACS</td>
<td>05-09</td>
<td>Number of people per square mile – calculated</td>
</tr>
<tr>
<td>Smoking (+)</td>
<td>BRFSS – Cnty Health Rank</td>
<td>2010</td>
<td>Percentage of population (&gt;18) who smoke</td>
</tr>
<tr>
<td>Alcohol (+)</td>
<td>BRFSS – Cnty Health Rank</td>
<td>2010</td>
<td>Percentage of population (&gt;18) who consume &gt; 5 (male) or 4 (female) alcoholic beverages at a time</td>
</tr>
<tr>
<td>Exercise (-)</td>
<td>BRFSS – Cnty Health Rank</td>
<td>2010</td>
<td>Percentage with less than daily recommended exercise</td>
</tr>
<tr>
<td>Indicator</td>
<td>Source</td>
<td>Year</td>
<td>Description</td>
</tr>
<tr>
<td>-----------</td>
<td>--------</td>
<td>------</td>
<td>-------------</td>
</tr>
<tr>
<td>Obesity (+)</td>
<td>BRFSS – Cnty Health Rank</td>
<td>2010</td>
<td>Percentage of population (&gt;20) with BMI &gt; 25</td>
</tr>
<tr>
<td>Mammmography Units (-)</td>
<td>FDA</td>
<td>2010</td>
<td>Number of certified mammography units per 1,000 women in county</td>
</tr>
<tr>
<td>“poor” general health (+)</td>
<td>BRFSS – Cnty Health Rank</td>
<td>2010</td>
<td>Percentage of population ranking health as “poor”</td>
</tr>
<tr>
<td>Low Birth Weight (+)</td>
<td>BRFSS – Cnty Health Rank</td>
<td>2010</td>
<td>Percentage of live births with babies weighing less than 5 pounds.</td>
</tr>
<tr>
<td>No social support (+)</td>
<td>BRFSS – Cnty Health Rank</td>
<td>2010</td>
<td>Percentage of population reporting no social support</td>
</tr>
<tr>
<td>Number of doctors (-)</td>
<td>Area Resource File</td>
<td>2010</td>
<td>Number of practicing doctors per 100,000 population</td>
</tr>
<tr>
<td>Number of internal MDs (-)</td>
<td>Area Resource File</td>
<td>2010</td>
<td>Number of Internal Medicine DRs per 1,000 population</td>
</tr>
<tr>
<td>Hospitals with oncology service (-)</td>
<td>Area Resource File</td>
<td>2010</td>
<td>Hospitals with oncology services per 1,000 population</td>
</tr>
<tr>
<td>Mammogram/pap smear &lt;2yrs (-)</td>
<td>BRFSS</td>
<td>2009</td>
<td>Percentage of population getting recommended screening in last 2 years</td>
</tr>
<tr>
<td>Uninsured population (+)</td>
<td>SAHIE</td>
<td>2010</td>
<td>Percentage of population without health insurance</td>
</tr>
</tbody>
</table>

### 3.2.1. Factor 1 - Social Characteristics

The contribution of these characteristics to the place based vulnerability model is based on their measurement of societal characteristics associated with higher cancer incidence or mortality in previous studies. This assumed connection is based on the ability of these indicators to capture the support systems and relationships within the community that may influence the ability to cope with health problems. There are 8 indicators in this factor, with the age being controlled through use of the age-adjusted incidence and mortality rates.

#### 3.2.1.1. Age/Gender (controlled variable)

Decline in health with age is a significant predictor of changes in vulnerability to negative cancer outcomes (Yang 2012). As an individual ages, the probability of both cancer incidence and mortality increases to a point, then incidence rates fall at further advanced ages (Frank 2004). There is also a difference in the rates tracked by age and gender. Male incidence and mortality rates tend to be higher through all ages and most sites (Cook 2011), but women tend to demonstrate lower rates until post-reproductive
years, when their rates increase. This suggests a hormonal component to cancer vulnerability, providing a protective benefit to women. The percentage of men in an area will therefore increase the probability of cancer incidence and mortality through all ages. There are also certain cancers, which are gender specific, or at least far more prevalent in one gender or the other. Two prominent examples are breast and prostate cancers, which account for a very high percentage of overall cancer incidence and mortality. Age and gender are controlled in this study through the use of age-adjusted, all-cancer rates.

3.2.1.2. Religious Affiliation and Marriage

Affiliation with a religion serves as proxy for social networking and support in the event of a cancer diagnosis. In a study done on breast cancer mortality, links to social networks such as marriages, club memberships, religious affiliations, and number of first-degree relatives, proved strong indicators of survivorship. Married women with high social support (group membership) tended to have lower mortality rates, while women with a high number of first-degree relatives and high social support tended to have higher mortality rates. Group membership alone did not have a significant impact on mortality (Kroenke et al. 2012).

For men, marriage has the effect of lowering cancer mortality rates across all groups. A study investigated the link between marriage and cancer screening, treatment and follow-up behaviors among men (Chamie et al. 2012). Married men had lower cancer mortality rates, which were linked to a greater likelihood for early diagnosis and follow-up treatments. Marriage also has the added benefit of increased income and
access to a spouse’s healthcare, especially for women (Koball et al. 2010). The marriage variable is percentage of the population over 18 now married (Table 3.1).

Membership in a religious congregation is used as a variable to represent the networking possibilities that could come from belonging to a religious congregation. Belonging to a religious group may influence the network of access to health care and treatment. The United States Religious census is used to ascertain the percentage of the county population belonging to any religious congregation (Grammich et al. 2012). Congregation membership from the US Religious Census is used as the variable (Table 3.1).

3.2.1.3. Self-reported Social Support

This indicator comes from the BRFSS survey and is collected from County Health Rankings data (Table 3.1). It is based on individual’s answers to a question about whether they feel as though they have adequate social support (University of Wisconsin 2010). Social support may not be relevant in the incidence of cancer, but it could be very important in the survival rates. People with greater social networks would likely be better equipped to cope with the stresses of cancer treatment, although the extent and type of cancer does play a role (Ell et al. 1992). The percentage of county population reporting low social support is potentially indicative of less connection, which could influence access to health care. Counties with higher social support may display a higher amount of connectivity amongst the population that leads to a higher likelihood getting support and following through with treatment.
3.2.1.4. Race/Ethnicity

The representation of different races and ethnic origins in a place serves as an indicator in this research for potential cancer vulnerability. There are a number of studies conducted on African American and Hispanic populations in the US related to breast and prostate cancer. Two studies have identified a far higher rate among African American men of prostate cancer mortality when compared to men of European descent (Brawley 2012; Taksler, Keating and Cutler 2012). This population has a 2.5 times higher risk of dying from the disease than the general male population. A study conducted on Hispanic and Black women also found a much higher mortality risk than among the general female population for breast cancer (Banegas and Li 2012).

For the purposes of this study, the race and ethnicity representation in each county will be measured as the percentage of the population that is not white (Table 3.1). According to the Census Bureau, the non-white population is the percentage of the county is not Black or African American, American Indian or Alaskan Native, Asian, Native American or other Pacific Islander, or some other race. The non-white population indicator is used in order to capture all minority races in a single variable. The American Cancer Society allows for multiple classifications of race, with rates measured for each separately. If multiple races and/or ethnicities were used in this research it would potentially over-represent this indicator in the analysis. The percentage of non-white population represents the likelihood of a county having ethnic minorities that may have difficulty accessing health care resources. While not an ideal way to identify ethnic enclaves on a smaller scale, it should still shed some light on counties with higher minority populations and possible associations to the MIR.
Genetics also play a very important role in certain cancer types, and specific ethnicities tend to have specific risk profiles. In research conducted by Sloan and associates in 2009, the geographic patterns of cancer incidence were compared with the distribution of genetic subpopulations. What they found was a relationship between the incidence of cancer and the presence of certain genetic subpopulations. For example, males of African American descent displayed far greater incidence of prostate cancer linked to genetic markers. Geographically, the distance from the centroid of the identified population correlated to the incidence rates. This demonstrates a geographic association between the genetic/ethnic background of populations and the incidence of cancer (Sloan et al. 2009). Unfortunately, the mapping of genetic markers is not prevalent across the country. When this data does become available it will likely add a great deal of information about the baseline vulnerability of populations.

While genetic predisposition may be a useful tool in future research, the theoretical basis for race and ethnicity inclusion in this study is based on the larger social constructs represented by racial/ethnic groups. The percentage of a county population belonging to an ethnic minority group provides the context to explain the MIR. There is a potential spatial component to this measure, with ethnic groups tending to cluster in space. Along with this spatial clustering, there is also an increased probability of social isolation as a result of social or language barriers that may reduce access to health. In addition, certain ethnic/racial groups may be more prone to exhibit other social characteristics like living in a single parent household, being unmarried, or renting.
3.2.1.5. Dependents and Single Parent Households

A transition to parenthood brings mixed benefits and drawbacks with respect to the health of the parents. Studies have shown both sides of the parental equation, largely depending on the age of becoming a parent. Parents with children, and to greater extent single parents, are typically busier and have less expendable income than their child-less counterparts. Children place a great deal of stress on relationships and the health of the parent. The same stresses also accompany the presence of elderly dependents in the household. In this research, households with dependents are considered those with children under eighteen or over sixty-five.

Adding single parent households to the analysis provides another means of gauging the lack of support and income that may be present in a community. In single parent families, the stress is greater due to the lack of support (Umberson, Prudrovskas, and Reczek 2010). For this reason, dependents in two-parent households, and to an even greater extent, single-parent households, are typically associated with poor health.

3.2.1.6. Language Isolation

Lack of communication is cited in a multitude of studies as having adverse impacts on both mental and physical health (Kang et al. 2010; Mora et al. 2010). Language isolation is related to social isolation, and typically presents as such, with people reporting depressive symptoms and higher levels of stress. The language isolation indicator in this study is collected by the census and accounts for households where no English is spoken. The percentage of non-English speaking households serves as proxy for language isolation and represents the portion of the population likely to face
communication challenges when seeking medical care. This could make a difference in health outcomes.

3.2.1.7. Renters

The percentage of renters in the county is a SoVI variable used to represent potentially low income or transient populations (Cutter, Boruff, and Shirley 2003). These groups are potentially vulnerable due to their lack of resources in the event of an emergency. The same vulnerability may apply to health, and these populations may lack the ability to cope or muster the resources required for adequate treatment. There is also the possibility that transient populations, represented as renters, may not have a general practitioner or other doctors (like OB/GYN) required for referrals or general health screenings (Roux 2010). The MIR may be influenced by this variable due to the barriers imposed from a combination of transience and lack of resources.

3.2.2. Factor 2 - Financial and Medical Access Characteristics

The ability to afford medical care, including screening services, routine check-ups, and treatment of illnesses, is crucial when considering cancer mortality. The likelihood of fatality is far greater when cancer is diagnosed at later stages or not treated quickly and effectively. Along the same lines, access to medical services can also be hindered through lack of access due to availability of services. Access to care can be divided into two sub-groups. Access can be defined as the ability to get care in a spatial sense as well as in a financial sense. Lack of access, regardless of the reason, leads to later cancer diagnosis and a lower likelihood of proper treatment for the disease (Wang 2012). Having the appropriate medical staff and facilities within easy reach may
influence health decisions. The number of facilities, number of qualified doctors, as well as the quality of the facilities and doctors must also be measured to determine the extent of spatial access (Shi et al. 2012). There are 9 indicators within this factor.

3.2.2.1. Access to Oncological Facilities

Two collected variables make up this indicator, mammogram units and hospital oncology units. Mammography units are certified by the FDA and counted by county as either stationary or mobile units. The presence of mobile units in a county is theoretically a much better option because populations can be targeted and it takes much of the hassle out of getting the procedure done. Many of the units go to areas of higher need, or will come to places of business to target vulnerable populations. Getting regular mammograms does have a strong correlation to catching breast cancer at an earlier stage and thus to a higher rate of survival (Helquist 2010).

Hospitals with oncology units play an important role in cancer treatment. There are tiers of cancer facilities, ranging from national centers all the way down to local centers. In this research, all hospitals with an oncology unit of any kind are counted for the county and normalized for the population. This is a rudimentary measurement and may not be the best to measure true proximity, but it is a good start at the national level. There are a multitude of reasons why close proximity to an oncology unit would be important for survival. One reason would be related to the presence of diagnostic and treatment capacities close by. Time is an important consideration in cancer cases, so both the stage of diagnosis and the time between diagnosis and treatment are many times critical. Having to travel long distances for treatment may delay this and lead to lower survivorship. The other reason proximity may be important relates to the support
network concept. If a patient must travel to another location for treatment, this may remove them from their support network or even reduce their likelihood of being treated at all. There is no evidence to support this assumption. There is a study, however; that demonstrates a relationship between survival and the distance to a cancer center (Lamont et al. 2003). Patients living at distances of greater than 15 miles from a treatment center were found to have an association with increases in survival of more than two-thirds compared to patients living within the 15-mile boundaries. This suggests that there are unmeasured variables influencing the outcomes. Either way, the impact of a cancer center in a place should demonstrate some connection to the cancer outcomes and will be measured in this research as the number of centers in a county per capita.

3.2.2.2. General Practitioners and Oncologists

The number of active general practitioners and oncologists as a function of the population may provide some insight into the likelihood of the population seeking regular care. Doctors are important in the maintenance of general health. Regular checkups and other routine vaccinations are critical in keeping good health and preventing future issues such as cancer (Sczcepura 2003). The variable for doctors is collected from the Area Resource File and normalized for the county population. Oncologists are important in both the diagnosis and the treatment of cancer cases. The presence of more oncologists in a county may provide more reliable detection of cancer at early stages and help lead to lower fatality rates (Ananthakrishnan 2010).
3.2.2.3. Insurance

Individual access to medical care can be measured as well as by possession of medical insurance. Individuals of lower income levels and without private insurance are at far greater risk of cancer mortality than those individuals with insurance and greater wealth (Koroukian, Bakaki, and Raghavan 2012; Robbins et al. 2010). This is attributable to the decreased likelihood of early detection through screening and proper treatment.

3.2.2.4. Median Household Income

Household income serves as proxy for both the ability to access screening, seek adequate preventive medical care and afford treatment. In the case of cancer, even with insurance the costs of treatment can overwhelm households without the financial means to pay for treatment. Identification of median household income may correlate with higher cancer fatality rates due to lack of appropriate treatment despite the presence of other positive variables, such as access to care (Cutter, Boruff, Shirley 2003; Tangka et al. 2010).

3.2.2.5. Income Inequality

The median household income of a county represents only a portion of the economic characteristics in a place. The GINI coefficient provides a method to measure the distribution of income among households in the county. The data collected for this research came from County Health Rankings, where values are standardized to fall in a range between 0 and 100, where 0 represents complete income equality and 100 represents complete inequality (University of Wisconsin 2010). A county may have a
median income that is artificially inflated by only a minority of the population. In this instance, one portion of the county may have a majority of the access to medical care. The GINI values are designed to indicate counties with large income disparities, possibly driving health disparities as well. These inequalities may represent a poor distribution of health resources and impact the health of the population.

3.2.2.6. Unemployment

Unemployment is added as a potential variable because of its connection both to the household income as well as to insurance coverage. Having a large unemployed population in a county could also indicate poor economic conditions leading to lack of adequate health care infrastructure. Unemployment at the family level could result in increased stress, loss of income and savings, as well as loss of health insurance. Any of these factors alone could increase vulnerability to negative health outcomes. Routine health checks and screening are less likely to occur among the uninsured population, and any extensive treatments may be forgone due to lack of insurance and income that accompany employment (Roetzheim 1999). The unemployment variable is from the Census and represents the percentage of the population over 25 that are unemployed.

3.2.2.7. Educational Attainment

Two major studies have related cancer incidence and mortality to the level of education an individual possesses. Chen et al. (2011) looked at the mortality rates for oral and pharynx cancer in relation to educational attainment, race/ethnicity, sex, and association with human papillomavirus (HPV) infection. The findings indicated a strong association between the cancer mortality rates and greater than a high school diploma.
The association is based on differences in the decreasing prevalence of smoking and sexual activity with higher education levels.

Another study examined the association between educational disparities and premature deaths from cancer. While the overall cancer mortality rate in the United States has actually been decreasing when the entire population is measured; however, among the populations with less than 12 years of education, the decrease has been much slower or even non-existent when compared with those having a higher level of education (Ma et al. 2012). In both studies, a high school diploma is used as a means of stratifying the study populations. Areas with a higher percentage of high school graduates are shown to have a lower risk of cancer mortality in each of the studies. Attainment of a high school diploma, for these reasons, will be used in this research as a proxy for education level. (Table 3.1).

3.2.3. Factor 3 - Community and Environment Characteristics

The space surrounding a community is important for both the physical and mental health of the community because it influences the perception of a community as well as the likelihood that social networks and outdoors activities will occur. The perception of a community is based on the cleanliness of the environment and the amenities available. Parks and other greenspace contribute to this perception and can make a place more inviting. On the same token there are characteristics related to the natural climate that may impact the ability or desire to get outdoors. Between the built and the natural environment, there is a fine line between having the convenience of urban life and the natural beauty of the rural. Each has advantages and disadvantages when it comes to the determination of health. There are 9 indicators within this factor that attempt to measure
the characteristics of a place related to the physical environment and access to both places for exercise and healthy food.

3.2.3.1. Population Growth

The population growth variable is calculated as the per capita change in county population between the decennial census of 2000 and 2010. (Table 3.1) There are a few major impacts that population change can have on a county. First, significant increases in population can overwhelm health infrastructure and create gaps in medical care due to lack of staffing (Galea 2005). Second, population loss can indicate economic challenges corresponding to unemployment, low income, or other social and environmental factors. Regardless of whether the population increases or decreases, significant changes in the population can have a deleterious impact on public health due to the strain placed on resources and the economy (Coleman and Rowthorn 2011).

3.2.3.2. Parks and Recreational Areas

The amount of greenspace in an area has been linked to the overall health outcomes of individuals residing in those areas. Two of the associated benefits, stress reduction and increased social cohesion, are proven to act as preventive measures against cancer incidence (Groenewegen et al. 2012). Greater amounts of greenspace, even grassy areas on roadsides, are associated with decreased stress levels and increased likelihood of active behaviors among residents. There have been other recent studies linking green space to lower cortisol levels, an indicator of stress and general wellbeing with increased greenspace close to the home and demonstrating a strong correlation to lower stress levels as well (Thompson et al. 2012).
Many studies have also linked the health benefits of exercise and a healthy diet with reduced cancer incidence (Block, Patterson and Subar 1992; Jew, AbuMweis and Jones 2009; Negri et al. 1991). Recreational facilities have shown to increase the likelihood of people engaging in physical activity, and along with parks and other public greenspaces, are indicators associated with increased social activities and improved social networks (Amodeo, Camera and Caimi 2010). The available variables come from the United States Department of Agriculture and the County Health Rankings. Recreational facilities data is available through a downloadable excel file from the FDA. Links and information on computation for this variable can be found in Appendix A. The number of parks is calculated per thousand in the county population and the recreational facilities are calculated per capita at the county level. In order to collect this data, the park layer was selected and downloaded from ESRI online data. This file contains 35,436 parks across the United States and is based on data collected in 2010 (Table 3.1). Using this data along with the population count for each county, the number of parks per 1,000 was calculated.

3.2.3.3. Natural Amenities Scale

This is an index created by the US Department of Agriculture to reflect the general livability of an area based on six measures of climate, topography, and water area that reflect environmental qualities most people prefer (USDA-ERS, 1999) (Table 3.1). The natural amenities scale was utilized in a North Carolina study that demonstrated a link between higher natural amenity scores, higher levels of physical activity, and BMI (Jilcott et al., 2011). There was found a positive association between the natural amenities scale and physical activity and a negative association with the BMI.
3.2.3.4. Environmental Hazards

The production and release of chemicals into the air and waterways can result in negative health outcomes for exposed populations. These releases come from vehicle exhaust, power generation, and the numerous companies that use or produce chemicals. In order to capture multiple aspects of the pollution in a county, two measures of air quality along with the total volume of releases are utilized in this research. Air quality measures are the number of days per year with unhealthy levels of ozone and the number of days per year where particulate matter is at an unhealthy level. Both of these measures have shown links to numerous health concerns, including cancer and are part of the clean air index (EPA 2009). The Toxic Release Inventory (TRI), maintained by the United States Environmental Protection Agency (EPA), provides chemical release information and includes all substances stored or released that pose harm to human health. The EPA, due to their known toxicity, regulates and tracks these substances. The EPA data is composed of county level counts of TRI emissions as raw counts as well as the count of days with unhealthy air. (Table 3.1)

3.2.3.5. Density and Rural Classification

The indicators for potential impacts of population density and the urban/rural divide are included in this research for a few reasons. First, there are positive impacts of residing in a rural area that could result from less exposure to pollution and more green space. Second, a downside could come from lack of access to adequate health care networks. In regards to the urban-rural impact, most research tends to agree on the fact that rural areas are generally associated with other income and racial indicators (Smith,
Humphreys, and Wilson 2008). The expectation is that a relatively high correlation will exist between the density and rural/urban indicators, but the relationship to the MIR is unknown. There are studies supporting both a correlation between rural areas and higher cancer fatality rates as well as those showing evidence of a link between higher cancer rates and the urban poor (Elliot et al. 2004; Freeman 2006). Two variables are collected to measure this characteristic. The census variable for percentage of a county designated as rural is used along with the population density (Figure 3.1).

3.2.3.6. Access to Healthy Food, Fast Food and Liquor Stores

Environmental access is measured through three separate indicators, the access to fast food, liquor stores and healthy food options. There are correlations between the level of access to both healthy foods, unhealthy foods and multiple correlates of health (Pearce, Blakely and Bartey 2007). Each of these indicators also shows a strong correlation to the SES of the neighborhoods as well, with a correspondence to “food deserts” (Beaulac, Kristjansson, and Cummins 2009). Lower income areas tend to have higher fast food and liquor store densities along with less access to healthy food options, as measured by travel distances. The number of fast food restaurants as well as the number of liquor stores in a neighborhood has shown in other research to correspond to poor health of the population (Block, Patterson and Subar 1992; Jew, AbuMweis and Jones 2009; Larson and Story 2009). The answer to whether the health problems or the stores and restaurants come first has not been answered. Despite this uncertainty, the correlation between the proximity of food options and the health of the individuals in the community is well enough established to merit the inclusion of these indictors. The connection between food deserts, liquor store density and cancer disparities still lacks much evidence,
although other health trends would indicate there should be some connection (Negri et al. 1991). Health food access is attained by calculating the percentage of zip codes in a county with healthy food options. Fast food access comes from the US Department of Agriculture and is calculated as the number of fast food restaurants per thousand in the population. The liquor store density also comes from the County Health Rankings and is derived using the number of liquor stores per square mile in a county. (Table 3.1)

3.2.4. Factor 4 – General Health and Behavioral Characteristics

The list includes high-risk behaviors such as alcohol and tobacco use; preventive behaviors like regular screening, and other general health status indicators including obesity, low birth weight and poor health. The 8 indicators in this factor are included because of their correlations with general health or with cancer directly, and are tied to behaviors and environmental exposures. Based on much of the research, many of these indicators were expected to display a strong correlation to cancer fatality rates, although no a priori assumptions of cardinality are implied. This also means there should be a relatively high amount of spatial variability among the data sets.

3.2.4.1. Smoking and Alcohol Consumption

Smoking is one of the more obvious indicators for cancer rates, and numerous studies into the carcinogenicity of tobacco products have demonstrated both a direct link to cancer as well as a correlation to SES (Carbone 1992; Hecht 2012). More recent research done by the Centers for Disease Control (CDC) reveals that, not only does smoking increase the risk of many cancers; it also prevents the body from effectively fighting it, potentially having significant impacts on the fatality (US-DHHS 2010). The
variable utilized comes from the County Health Rankings and represents the percentage of the population over the age of 18 who smoke regularly. (Table 3.1) There are also studies linking cancer incidence and mortality to individuals who consume alcohol (Rehm et al. 2010). The cancers typically associated with alcohol intake involve the mouth, throat and digestive tract. In both indicators, there is a strong link to the SES of the population. Income and education have demonstrated a convincing link to higher risk behaviors such as smoking and alcohol (Hiscock et al. 2012; Huckle 2010). The alcohol indicator in this study is related to binge drinking, and is a percentage in each county of adults who have engaged in binge drinking in the last thirty days. According to the BRFSS, binge drinking is consuming more than five alcoholic beverages at a time for a male and four or more for a female. This variable comes from the County Health Rankings. (Table 3.1)

3.2.4.2. High Risk Occupations

Certain occupations have a far greater risk of cancer incidence and mortality due to exposures on the job, with farmers and blue-collar workers having the highest risks. Farmers exhibit higher risks of incidence primarily due to sun exposure and the risks associated with chemical exposures from pesticides and fertilizers. Blue-collar workers in the construction and hotel/catering exhibit higher cancer incidence and mortality rates associated with chemical exposure and inhalation of carcinogenic substances (Bouchardy et al. 2002). This data should prove comparable to that found in the US, with similar employment sectors and exposure in the field. The data for this variable comes from the 2009 Economic Census and is the percentage of the county population having an identified high-risk occupation as determined through NAICS codes. (Table 3.1)
3.2.4.3. Exercise

There are a multitude of studies relating exercise to health outcomes (Friedenreich and Orenstein 2002). Of the cancers investigated, breast and colon cancer incidence are shown to have a convincing link with exercise. Increasing amounts of physical activity have a preventive impact. In addition, there is mounting evidence of the preventive influence exercise exerts on prostate cancer as well as lung and endometrial cancers. The recommendations that have come from these studies consider acceptable amounts of physical activity to be thirty minutes of moderate to intense exercise at least five days a week. This is also found to influence cancer detection and coping mechanisms, which may impact the likelihood of survival amongst those diagnosed with all cancers. The link exercise has with cancer in this study is related to survivorship. The MIR measures the fatality of the disease, and exercise has shown a correlation to the health of individuals after diagnosis (Grimmett 2011). The collected variable comes from the County Health Rankings and is the percentage of the population getting less than the recommended thirty minutes of daily exercise. (Table 3.1)

3.2.4.4. Mammography Screening Behavior

Screening for certain cancers; breast, prostate, cervical and ovarian as prime examples, has shown varying connections with cancer prognosis. The benefit of screening tests is the ability to catch the cancer at an earlier stage and providing a better chance of survival with treatment. Mammograms are perhaps the best screening tests when it comes to early detection of cancer and improvement of prognosis. They are also some of the most routinely done of the cancer screening tests. The ACS completed a
survey of the association between screening behaviors and the five-year survival rate for all cancers. They found that the five-year survival rate was 70 – 90% higher among those identified through screening programs (ACS 2012). The variable collected comes from the County Health Rankings data and is calculated as the percentage of women over 40 who have gotten a mammogram in the past two years. (Table 3.1) It would have been useful to also assess the screening rates for colorectal and prostate cancer as well, however there is not a reliable source of county level data available for either of these variables. It would be beneficial to utilize other screening tests in future analysis, such as prostate-specific antigen (PSA) testing or colonoscopies. At the current time, however, data for these screening behaviors is not prevalent enough to use at the national level for all counties.

3.2.4.5. Obesity

Obesity, as measured by the Body Mass Index (BMI), has been linked to a number of cancers (Calle and Thun 2004). Colon, breast (postmenopausal), endometrial, and esophageal cancers have all proven to have causal links to obesity. Prevention of obesity is of paramount importance, as there is little evidence of successful long-term weight loss among individuals classified as obese. In England, 5% of cancer cases in postmenopausal women were attributable to being overweight or obese (Reeves et al. 2007). This particular study is focused on the increasing obesity rates in the country, considering this is an avoidable risk and could make a significant impact on cancer rates where 23% of women are classified as obese and 34% as overweight. In the US, similar trends in BMI are likely causing increasing prevalence of preventable cancers. This
variable also comes from the County Health Rankings and is the percentage of the population over twenty with a BMI above twenty-five. (Table 3.1)

3.2.4.6. Low Birth Weight

Low birth weight is included in this analysis due to its correlation a multitude of health disorders, including heart disease, hypertension, and diabetes later in life. The association seems to be independent of sex, ethnicity or health behaviors such as smoking or alcohol abuse (Valdez et al. 1994). The use of this indicator is based on research that demonstrates a link between the birth weights of babies and the health of the mothers. There are strong ties between this indicator and other population health indicators such as dietary intake, smoking, alcohol consumption, and stress levels (Kramer 1987). The definition of low birth weight is established by the Centers for Disease Control (CDC) as a baby weighing less than 5.5 pounds at birth. Data for this variable are collected from the County Health Rankings as the percentage of live births reported as low birth weights. (Table 3.1) Having a higher percentage of low birth weight babies in a county would be expected to exhibit lower population health for a county.

3.2.4.7. Self-reported poor health

This is a variable collected from the County Health Rankings and is derived from the BRFSS telephone survey. Respondents were asked whether they would rate their health as “good”, “fair” or “poor”. The relevance of this question in this research is the gauge of mental health. Despite other physical maladies, people with a good outlook toward their health are more likely to experience good outcomes when dealing with cancer (Sanne et al. 2012). This should impact the survivorship and lead to a lower MIR
in counties reporting better health. The variable collected comes from the County Health Rankings and represents the percentage of the population in the county ranking their health as “poor”. (Table 3.1)

3.3. Cancer Outcome Indicator - Mortality to Incidence Ratio (MIR)

The analysis in this research is conducted using the incidence and mortality rates for all cancers combined. Incidence and mortality rates for cancer each represent different issues and could be explored separately. For the purposes of this research, only the MIR is utilized as an outcome measure. Future research may dictate a more in-depth analysis of the incidence and mortality rates on an individual basis at a sub-regional level. Incidence and mortality rates, if measured on their own, would show potentially different relationships to the indicators.

The MIR is calculated using all-cancer incidence and mortality rates. This is done for two reasons. First, using all cancer rates allows for the inclusion of more counties and a more extensive analysis. Breaking out individual cancer types would likely reveal some different patterns and is definitely of interest for future research, but there are many counties that would have to be excluded due to too few cases. Second, the purpose of this research is directed at the identification of spatial cancer vulnerability. Using a multitude of different cancers would muddy the waters, making regionalization potentially difficult to discern. It is likely that more prevalent cancers such as breast and prostate dominate the data, however, the type of cancer affecting an area is of little interest in this research. A death from cancer is still a death, regardless of cancer type. By using the methods proposed in this research, the drivers of vulnerability are tested.
independently and should reflect the characteristics making a population vulnerable, regardless of cancer type prevalence.

The geographic distribution of the MIR is a very significant part of this research. The MIR is chosen as an indicator of cancer vulnerability due to the concept it represents, the deviation of expected cancer deaths based on the prevalence rates. A national average MIR specifies the chance of survival in a population with a defined number of cancer cases. Deviations from this average indicate either differences in treatment and detection or differences in cancer types. Vulnerable populations can be identified geographically by mapping the populations based on these deviations from the mean MIR.

The MIR is calculated using information from the National Program of Cancer Registries (NPCR) and reported as the ratio of the two indicators. (Table 3.1) Incidence rates and mortality rates were collected by state and checked for confidence levels. For each rate, the CDC reports at a 95% confidence rate based on a population of 100,000. The rates are age-standardized based on 5-year age groups to the 2000 U.S. standard million population. The CDC removes county level incidence and mortality data where there are three or less cases annually over the course of data collection. As discussed previously, the entire state of Kansas and Minnesota are not collected due to state laws that repress the data. Aside from that, all county data is available for the 2005-2009-collection term. Yakutat County in Alaska and Kalawao County in Hawaii are each repressed due to low incidence counts. Alaska and Hawaii are removed due to lack of availability for other indicators, however. In Montana, Petroleum County is removed. Arthur and Bradford Counties are removed from Nebraska, while Texas has Kenedy,
King, Loving, McMullen, Roberts, Sterling and Terrell Counties removed from the analysis due to incidence rates that are too low.

Also of interest for future analysis is the increased number of states and counties with repressed data. The data used for this research came from 2005-2009 estimates, while the new data available on the NPCR is 2006-2010. In this new data set, the states of Arkansas, Ohio, and Virginia now have no data available. In addition, Washington State now only has thirteen of thirty-nine counties with available data. Some of this data, Virginia and Ohio for example, can still be found on the state cancer registry sites. The other counties in Washington State are also available on the state registry website, just not through the CDC.

3.4. Summary and Model Operationalization

The indicators listed are intended to represent a multitude of societal factors contributing to the health of populations. There are two primary known facts going into this research, some coming from the field of epidemiology and some from hazards geography. First, each of the indicators has shown to possess a link to health or cancer. Second, there is a degree of spatial variability in both the cancer outcome and in the indicators.

Based on this knowledge, the goal of this research is to analyze both the extent of the spatial variability among the indicators as well the contribution of each indicator to the MIR outcome. The purpose of the conceptual model is to place each of the indicators into a role depending on the influence each indicator has shown to health or cancer outcomes in previous research. Grouping the indicators is intended to make the identification of themes more apparent. Many of these indicators measure similar
societal constructs and most likely have similar impacts on health. Many studies control for specific variables or eliminate highly correlated variables and could be missing important relationships in the process. In addition, the four theoretically derived factors are expected to display spatial patterns and potentially lead to a more complete understanding of spatial variations in the MIR. Comparison of patterns may reveal some of the relationships between the societal constructs and the MIR that enhance our ability to identify methods to eliminate the disparities. The conceptual model in this research allows for a more structured approach when assessing the impact of these variable interactions prior to testing their correlation with the MIR.

Operationalization of the conceptual model will first be accomplished by testing the individual indicators with the MIR to determine the correlation each one has with the MIR in this data set. Knowing the individual relationships will provide a baseline for analysis after the grouping of variables. The second step in operationalizing the model is to test the a priori groupings. The way in which these indicators are grouped is based on societal structures and assumptions as to how they will associate with each other and with the MIR. A Principle Components Analysis (PCA) is used to determine the groupings of the variables as well as the indicators with the most variance. The goal of the model is to determine cancer disparities, so finding data with the biggest divergence of values is important. The final test of the model involves testing the spatial relationship of the chosen indicators and groupings. This portion is where the model can provide useful information about both the indicators and the factor groupings as they vary through space. Regional trends will be determined through this test, as each factor is mapped and compared to the MIR. The factors within the model having the highest spatial correlation
to the MIR provide an insight into sub-regional patterns that merit further investigation as well.
CHAPTER 4

METHODOLOGY

This section will provide information on the study region, data collection, and the methods used to answer the research questions. There are three research questions in this study, each being represented by a separate chapter. The methods, results and analysis in each chapter are intended to answer each question independently before summarizing the final outcomes.

4.1. Study Region

The data collected is for the entire nation at the county level, permitting analysis of the many cancer trends that play out on a regional basis. For instance, cancer incidence and mortality rates are highest in the southeastern United States for most cancer sites (ACS 2012). The county is also the smallest enumeration unit available for some of the data sources. The census products have the capability to provide a finer spatial resolution, but many of the behavioral indicators, such as smoking, alcohol use, screening and environmental access are not available at smaller scales. More importantly, cancer rates are not available at a sub-county level for the entire United States. The all-cancer incidence rates are likely to be high enough in most counties, but mortality rates would not be high enough to achieve reliability in a large number of the rural counties at a smaller level. If a sub-county scale were launched prior to any large-scale analysis, there would be no point of comparison between the study area and other
places in the county. By starting with the entire United States first, not only can the regional trends be established, but also any subsequent research at a smaller scale can be compared back to the entire country. Regional trends can also provide insight into important characteristics that merit a smaller scale analysis, thereby not wasting precious resources where disparities do not exist.

4.2. Pre-processing of Data

Prior to any testing or data conversion, the data values are checked for gaps in either the variables or geographies. Indicators or geographies with too many missing values are removed to maintain the consistency of tests. There are 3,143 counties in the United States as of the 2010 census. In order to maintain statistical significance for the aspatial tests, there only needs to be data for 550 counties within each indicator for a 99% confidence level. There are no collected variables in the set that have less than the required number of values, therefore none require removal from the analysis. There are three variables dealing with environmental characteristics that must be removed, however. The natural amenities scale contains categorical data that cannot be accommodated with the remainder of the continuous variables. This data set is also highly compressed due to the assignment of categories and lacks variability when compared to other indicators. Ozone days and Particulate Matter days are also removed due in part to their categorical data distribution, in addition to having and excessive number of zero values that cause the data set to be highly skewed.

Geographically, the exclusion of Alaska and Broomfield County, CO are determined necessary based on the number of data points missing. Alaska is lacking approximately ten of the thirty-seven variables and has zero values for an additional three
variables. Also, the MIR is calculated at the state level and applied to each county, giving the state no variance for the outcome measure. Hawaii has a very similar issue with sparseness of data and is removed as well. Broomfield County, CO did not exist at the time of many of these data measurements, so also lacks many data points. Kansas and Minnesota also must be excluded from the analysis due to their lack of reporting cancer data. Indicator measures for these two states are still collected, but they were not included in the regression analysis or final results. The state averages for Kansas and Minnesota were tested in the hot spot analysis to see if the results came out different. The result of this test showed little difference from the analysis using no data in Kansas and Minnesota. The final data set contained 34 variables and values for 2868 counties in 46 states.

4.3. Procedures

This research had three main procedures intended to answer the primary research questions. These three analyses included a theoretical, inductive, and deductive approach. The key focus was on the relationship both amongst the variables and with the MIR. This analysis was not meant to be predictive so much as it was meant to flesh out the relationships existing among characteristics of populations that lead to higher cancer fatality. The goal was to identify the characteristics having the most contribution to cancer fatality in a place as measured by the outcome variable, MIR. In order to accomplish this, there are three different sets of analyses carried out.

The flow diagram in Figure 4.1 provides an overview of the three main steps conducted to examine the relationship of both the individual variables and the two factor grouping methods with the MIR. A description of each method follows.
4.3.1. Step 1 - Primary Indicators

The first procedure uses a deductive approach to compile the sociospatial indicators driving vulnerability to cancer deaths found in research questions one and two. Potential cancer vulnerability factors are collected at the county level and normalized for population. They are then standardized and analyzed for linearity and normality. In order to ensure a consistent positive relationship for all indicators to the MIR, the inverse is calculated for variables with negative correlations.

![Flow diagram showing the aspatial validation procedure for testing single factors versus multidimensional index.](image)

Running an initial regression using all thirty-four variables establishes a baseline for both the predictive ability of the entire set of variables as well as the contribution of each individual variable to the outcome measure (MIR). A multiple regression model is
run in SPSS to set up this baseline and begin to assess the influence of each indicator on the MIR. A correlation analysis is also conducted to determine relationships amongst the indicators. In the regression model, variance inflation factors (VIFs) are assessed for each indicator to get a better idea of the multicollinearity within the data set and beta values are evaluated to determine the relationship of indicators to the MIR. Any indicators with VIFs of greater than 2.5 along with tolerances of less than 0.4 are investigated further in a correlation matrix to determine their relationships with other variables in the data set. The rationale behind the cutoffs is based on expert opinion and previous research into multicollinearity (Allison 1998; Besley 1991; Wheeler 2010). Common guidelines dictate that a VIF greater than 10 should be considered for multicollinearity. Other researchers, think a much lower VIF of 2.5, along with a tolerance of 0.4, should be investigated more closely (Allison 1998). In the case of this research, the recommendations of Allison are followed due to the fact that this is social research and the model is not exceptionally strong. A VIF of greater than 2.5 could indicate multicollinearity if correlated highly with another single variable.

The goal of investigating the bivariate correlations is to ensure that the correlations between the high VIF variables are not with other high VIF variables. Highly correlated variables will cause a distortion of the model and over represent the constructs they characterize. The goal is not to eject any variables from the model unless they contribute to the distortion. The presence of correlated variables can shed some light on societal patterns that tend to vary together in space. Multiple correlates could, in fact, magnify the impact of each other and produce more . For this reason, those indicators with a high VIF will not be removed from the regression model. After analysis of the
VIFs, six variables are flagged for and checked for correlation to other variables. The six
variables are non-white population, median household income, education level, no
exercise, doctors per thousand in population, and renters. Each of these variables has a
high VIF and low tolerance. After looking at correlations, however, each of these
variables is retained in the model. No further multivariate regressions are run, and thirty-
four variables will be utilized in the grouping of indicators that follows in step two and
the spatial analysis conducted in step three.

4.3.2. Step 2 – Grouping of Indicators

The second procedure is intended to develop a more robust and consistent set of
indicators of cancer vulnerability. Two different approaches to groups were done: 1) an
a priori theoretical categorization based on the extant literature and 2) an inductive
classification based on principal components analysis making no assumption about
groupings.

4.3.2.1. Theoretical Grouping

Current indicators identified in the extant research were grouped according to a
priori assumptions of interactions between each variable, resulting in four distinct factors
identified in each of the health disparities models identified by Diez Roux (2012): social
characteristics, general health and behaviors, medical and financial access, and the
community and environment characteristics.

For the social characteristics factor, there are two main criteria for inclusion. The
first is based on social exclusion. Groups that are isolated from the mainstream
population will potentially be more vulnerable to negative health outcomes. The other
criterion involves support networks and potential to cope with stress. Individuals with strong support networks and less stress may be more likely to receive an earlier diagnosis due to screening, more likely to receive treatment, and more likely to survive. The conceptual basis for this grouping is derived primarily from the pathways model in disparities research. The pathways model is based on a combination of social constructs and the underlying ethnicity of the population. Ethnicity is measured by both the percentage of non-white and non-English speaking residents in a county. Social constructs are measured through marriage rates, household structures, religious adherence, and potential support network.

Measures of general health and behaviors are intended to capture the societal traits considered integral in many iterations of the pathways model. A bulk of the indicators in this factor can be classified as behavioral, including smoking, alcohol use, exercise and screening. These represent personal, individual choices, but are driven in large part by access issues and other social variables existing at a community scale. The distal causes represented by this grouping may point to potential mediating factors considered important by the pathways model due to their modifiable nature and importance in eliminating existing disparities.

The physical and financial access indicators are chosen in large part based on the fundamental cause model of health disparities. The primary basis of this grouping lies in the economic influence of the population. Higher household incomes will secure better health through other characteristics such as education and better access to facilities and doctors. The indicators utilized, in addition to the economic, include access variables such as number of doctors, internists, oncology facilities, and screening units. Higher
income populations are also expected to correspond to higher income inequality, higher education levels, lower unemployment and higher number with health insurance.

The community environment indicators are intended to represent neighborhood level characteristics that may influence the health. Variables such as parks and recreational facilities can reveal information about both the ability of residents in a county to access places to exercise and socialize as well as the importance placed on these amenities by the community. The presence of healthy versus unhealthy food options should also drive community health, with a higher density of liquor stores and fast food restaurants most likely linked to poorer health. The variable measuring the percentage of renters in the community is intended to address overall livability and potential transience. Higher numbers of renters indicates a poor housing market and lower community identity. The population is not tied to the area in the same way that homeowners would be, and this could manifest in a less appealing place. The conceptual foundation of this factor is based on the fundamental cause model, with the assumption that many of the indicators will be linked to the resource base of the community. The SES of the county would drive allocation of resources that benefit health, while minimizing the placement of entities causing adverse health.

To test the reliability of the theoretical groupings, a Cronbach’s Alpha test is utilized. The consistency of the theoretically created factors is also tested against the PCA factors in relation to the outcome. Analyzing the two different grouping methods, theoretical versus PCA, sheds light on some additional qualities of the variables and how they may influence the cancer outcomes.
4.3.2.2. Inductive Grouping

In addition to the theoretical grouping method, an inductive method was also used, and involved entering the identified cancer vulnerability indicators, listed in Table 3.1, into a PCA. The PCA method was used to statistically reduce variables into categories that explain a majority of the variation between the counties. PCA helps to identify variables that tend to correlate with each other and group them into orthogonal factors and also reveals pairing of indicators not previously considered. This method is similar to that used in the Social Vulnerability Index (SoVI) (Cutter, Boruff and Shirley 2003), but enhanced to account for drivers of negative health outcomes. Optimized factors were created by using a varimax rotation and retained if they met Kaiser Criterion, an eigenvalue of greater than one. This ensured factors that were orthogonal and explained by more than one variable in the set. By eliminating the indicators with little variation between the counties, a more significant and consolidated set emerges to provide a clearer picture of the variables influencing cancer fatality. The goal of this analysis was to determine where the greatest possibility exists for disparities in the MIR amongst US counties. The loading plots for the factor analysis are also analyzed to assess the potential constructs measured by each factor.

4.3.2.3. Analytical Procedures

There are two methods by which the groupings were analyzed. The first involved a regression analysis and was intended to examine the relationship of the factors to the MIR. The regression analysis first implemented a comparison of predictive ability between the individual model and the two grouping methods. This was an important
step, and revealed loss of explanatory power incurred through the grouping of indicators. The goal is to identify the most concrete set of indicators possible and prove that the groupings are representative of real world patterns. In addition to testing the external consistency of the factor groupings as they relate to the MIR, the regression also unveils some evidence of relative influence of each factor to the MIR. This is determined through analysis of the beta-values for each factor created through either theoretical or inductive grouped indicators.

Predictive powers of both the summed PCA factors as well as the independent regression variables are compared to determine the influence of variable reduction. The methodology described is outlined in Figure 4.1. Some loss is expected due to the lesser number of independent variables; however, the benefit of the method lies in the reduction of this multitude of variables into a more comprehensible set. In doing so, some of the relationships between variables may be highlighted and shed light on how they play out in relationship to one another.

To test the resulting factors of the PCA output, a regression model is run using the ten factors identified. This is conducted in the same manner as the regression with all variables, and the adjusted $R^2$-value will reflect the predictive value of the set. The factor scores are derived using the transformed factor values in SPSS for each indicator being used. This method creates eight non-correlated factors constructed based on the weights of the variable contributions. These new representative variables can be tested in the regression model. The adjusted $R^2$ is analyzed for this model in order to accommodate the different numbers of variables between the initial multiple regression and the post-
PCA multiple regression. In addition, the significance of each derived factor coming from
the PCA model can be tested in relation to the MIR by assessing the beta values.

The concept of the factor approach is two-sided, based on both the potential for
characteristics of a place to feed off one another as well as the ability of the factor to shed
light on spatial constructs influencing health. The interactions amongst variables may be
either antagonistic or synergistic, depending on what they measure and how they interact
with each other. For example, low-income levels may have a more pronounced impact on
the MIR than expected when considered in conjunction with single-family households,
indicating a synergistic relationship. If there are synergistic interactions between
variables, the groupings should reveal increased $R^2$-values when compared to the
individual variable regression. Correlation matrices can also be checked to determine the
influence of grouping the variables by the two different methods.

Final analysis of the inductive approach involves a path analysis using a
combination of the a priori group identifications and PCA grouping of variance data. The
multiple paths are checked using SPSS-AMOS software to calculate correlations and
regression coefficients between all of the groups. Structural equation modeling is then
utilized to test paths and find the combination of factors and order to yield the highest
total correlation with the MIR. This data is used to compare with the conceptual model
and make any required adjustments.

4.4. Step 3 – Spatial Variability of Factors

The methods explained above are aspatial, seeking to identify the more influential
indicators of cancer vulnerability and disparities without consideration for patterns
between places. The decision to incorporate an analysis of spatial variation amongst the
indicators is primarily policy-based. Identification of significant relationships between vulnerability drivers and the MIR may help to identify locations in the U.S where specific factors lead to higher or lower cancer fatalities. The vulnerability profile of a county will be the significant outcome of this research, and will give policy-makers good insight into the factor(s) driving the vulnerability.

The third research question, in search of regional differences in the data, requires spatial techniques and is answered through the use of a hot spot analysis conducted on the significant factors identified from both the theoretically grouped factors and the inductive grouped factors as well as on the MIR outcome measure. Vulnerability factors identified in the theoretical and inductive methods are initially assessed for spatial clustering along with the MIR. The goal is to identify locations with either vulnerability factors or MIR varying significantly from the mean. Hot spots are clusters with higher than average incidence, while cold spots exhibit lower than average incidence.

To obtain the hot and cold spot locations, a Global Moran’s I analysis is utilized to determine the extent of spatial clustering present in the data. This serves as a pattern detector, and identifies results as clusters, dispersed, or random arrangements of values. In this analysis, the conceptualization of distance is based on a fixed distance band measured by Euclidean distance. If evidence of clustering exists, a Getis-Ords Gi* statistic identifies and measures the type of correlation that exists between the locations. High or low values clustering together are identified through the use of these two methods, together referred to as a hot spot analysis. The goal is to identify counties demonstrating a clustered pattern for MIRs of higher or lower values and assess the corresponding clusters of vulnerability factors as they compare to the MIR. Visualization
of the patterns also helps to identify patterns that may not be evident when only investigating the aspatial relationships.

Clusters can also reveal regional or state level patterns related to political borders or other policy influences. The relationship of MIR clusters to the vulnerability factor clusters may also expose details pertaining to the blend of characteristics present in each county.

In addition to the visual comparison of the hot and cold spots, a correlation analysis is also run using the Gi* values for the MIR as well as the factors to counties derived from the analysis. The objective is to identify spatial correlations existing in the data, with results of the model compared to both the aspatial regression models as well as to the visual pattern results from the hot spots analysis.
CHAPTER 5

PREDOMINANT VULNERABILITY INDICATORS

This chapter addresses the first research question involving the identification of predominant socio-spatial variables driving vulnerability to cancer deaths in the US. The analysis begins by investigating the connection between MIR and the set of initial indicators. Following pre-processing of data, an initial regression analysis is conducted using the thirty-four indicators to serve as predictors of the MIR. The intent of the model is to establish a baseline for the predictive ability of the entire data set as well as to determine the relationship of each variable to the MIR.

5.1. Predictive Ability and Model Adequacy

The results of this first regression, shown below in Table 5.1, show an adjusted $R^2$ for the entire set of 0.347. The F-statistic and the Durbin-Watson (d) statistic are also calculated to establish the fit of the model.

Table 5.1. Regression results for 34-variable model.

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Standard Error of the Estimate</th>
<th>Durbin-Watson</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.596$^*$</td>
<td>.355</td>
<td>.347</td>
<td>.791720701500</td>
<td>2.012</td>
</tr>
</tbody>
</table>

*Note: $R^2$ is the coefficient of determination, representing the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.
The ANOVA table results in an F-statistic of 46.346 with a p-value of 0.00, indicating a good model fit. The d-statistic is 2.012, indicating independence of the variables in the data set. The coefficients (beta values) and variable inflation factors (VIFs) in the model, shown in Table 5.2, are also checked to ensure that no multicollinearity exists in the set.

There are a total of eleven variables with a VIF of greater than two, including non-white, unmarried, single parent households, renters, income levels, education levels, doctors, internists, rural population, obese population, and people not getting adequate exercise. Each of these variables is analyzed in a correlation matrix to determine what drives the higher VIFs. None of the correlations are higher than 0.745. In addition, the variables with higher VIFs tend to exist in groups. For example, in one group there is the percentage of non-white population correlates with the unmarried population (0.667), number of single parent households (0.745), and the number of low birth weight babies born (0.611). Another group is tied together by education level, which correlates with income level (0.698), percentage of people not getting enough exercise (0.651), number of doctors (0.566), obese population (0.554), number of internists (0.552), rural population (0.550), and the number of women getting mammograms (0.514).

Based on the correlation analysis, the variables representing race/ethnicity and education level are responsible for the higher VIF values. The two highest VIFs are education level, at 5.713 and the percentage of non-white residents, at 5.411. (Table 5.1) The decision is made not to remove any variables, however.
Table 5.2. Regression model beta values and collinearity statistics for 34 variables. MIR is dependent.

<table>
<thead>
<tr>
<th>Model</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
<th>Collinearity Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Beta</td>
<td></td>
<td></td>
<td>Tolerance</td>
</tr>
<tr>
<td>Non-white</td>
<td>.033</td>
<td>.914</td>
<td>.361</td>
<td>.185</td>
</tr>
<tr>
<td>Religious Adherence</td>
<td>.075</td>
<td>4.083</td>
<td>.000</td>
<td>.711</td>
</tr>
<tr>
<td>No social support</td>
<td>.005</td>
<td>.245</td>
<td>.807</td>
<td>.678</td>
</tr>
<tr>
<td>Unmarried</td>
<td>.091</td>
<td>3.172</td>
<td>.002</td>
<td>.290</td>
</tr>
<tr>
<td>Single parents</td>
<td>.097</td>
<td>3.442</td>
<td>.001</td>
<td>.301</td>
</tr>
<tr>
<td>Language Isolation</td>
<td>-.019</td>
<td>-1.194</td>
<td>.233</td>
<td>.906</td>
</tr>
<tr>
<td>Renters</td>
<td>-.009</td>
<td>-3.37</td>
<td>.736</td>
<td>.364</td>
</tr>
<tr>
<td>Dependents</td>
<td>.008</td>
<td>.355</td>
<td>.722</td>
<td>.478</td>
</tr>
<tr>
<td>Median Household Income</td>
<td>.004</td>
<td>.114</td>
<td>.910</td>
<td>.196</td>
</tr>
<tr>
<td>GINI</td>
<td>.049</td>
<td>2.064</td>
<td>.039</td>
<td>.425</td>
</tr>
<tr>
<td>Education level</td>
<td>.081</td>
<td>2.201</td>
<td>.028</td>
<td>.175</td>
</tr>
<tr>
<td>Unemployment</td>
<td>.013</td>
<td>.665</td>
<td>.506</td>
<td>.592</td>
</tr>
<tr>
<td>Mammography Facilities</td>
<td>.012</td>
<td>.678</td>
<td>.498</td>
<td>.745</td>
</tr>
<tr>
<td>Oncology Facilities</td>
<td>.014</td>
<td>.786</td>
<td>.432</td>
<td>.806</td>
</tr>
<tr>
<td>Uninsured</td>
<td>-.008</td>
<td>-3.56</td>
<td>.722</td>
<td>.431</td>
</tr>
<tr>
<td>Doctors</td>
<td>.008</td>
<td>.321</td>
<td>.748</td>
<td>.357</td>
</tr>
<tr>
<td>Oncologists</td>
<td>.035</td>
<td>1.420</td>
<td>.156</td>
<td>.392</td>
</tr>
<tr>
<td>Parks</td>
<td>-.025</td>
<td>-1.509</td>
<td>.131</td>
<td>.901</td>
</tr>
<tr>
<td>Rec Centers</td>
<td>-.018</td>
<td>-9.88</td>
<td>.323</td>
<td>.760</td>
</tr>
<tr>
<td>Rural</td>
<td>.115</td>
<td>4.309</td>
<td>.000</td>
<td>.334</td>
</tr>
<tr>
<td>Pop Growth</td>
<td>.016</td>
<td>.975</td>
<td>.329</td>
<td>.869</td>
</tr>
<tr>
<td>Density</td>
<td>.006</td>
<td>.377</td>
<td>.706</td>
<td>.973</td>
</tr>
<tr>
<td>Liquor stores</td>
<td>-.030</td>
<td>-1.760</td>
<td>.079</td>
<td>.825</td>
</tr>
<tr>
<td>Fast Food</td>
<td>.031</td>
<td>1.495</td>
<td>.135</td>
<td>.574</td>
</tr>
<tr>
<td>Health Food</td>
<td>-.049</td>
<td>-2.963</td>
<td>.003</td>
<td>.885</td>
</tr>
<tr>
<td>Obese</td>
<td>.176</td>
<td>6.409</td>
<td>.000</td>
<td>.318</td>
</tr>
<tr>
<td>Poor Health</td>
<td>.006</td>
<td>.323</td>
<td>.747</td>
<td>.622</td>
</tr>
<tr>
<td>TRI Release</td>
<td>-.021</td>
<td>-1.322</td>
<td>.186</td>
<td>.974</td>
</tr>
<tr>
<td>Low Birth Weight</td>
<td>.094</td>
<td>3.854</td>
<td>.000</td>
<td>.402</td>
</tr>
<tr>
<td>High Risk Occupations</td>
<td>-.006</td>
<td>-3.22</td>
<td>.747</td>
<td>.754</td>
</tr>
<tr>
<td>Smokers</td>
<td>.079</td>
<td>3.978</td>
<td>.000</td>
<td>.602</td>
</tr>
<tr>
<td>Binge Drinking</td>
<td>-.041</td>
<td>-2.051</td>
<td>.040</td>
<td>.607</td>
</tr>
<tr>
<td>Lack of exercise</td>
<td>.068</td>
<td>2.399</td>
<td>.017</td>
<td>.301</td>
</tr>
<tr>
<td>Regular Mammogram</td>
<td>.053</td>
<td>2.592</td>
<td>.010</td>
<td>.568</td>
</tr>
</tbody>
</table>
Considering the model parameters that demonstrate significance, the fact that none of the correlations are above 0.8, and that the two major variables influencing any multicollinearity are important to the mode, there is no benefit to eliminating any of these indicators. To confirm that no model improvement is evident, the education and race/ethnicity variables were removed to run another regression model. The adjusted $R^2$ goes remains at a 0.347, indicating that no model improvement occurred as a result of the removal. Taking all of this information into account, the 34 variable regression model does appear to be adequate for predictive purposes and will serve as an adequate baseline comparison to the grouped models. The full variable set is used to group and create the factors for further testing of the conceptual model.

5.2. Identifying Significant Indicators

The other objective of this initial regression analysis was to determine the level of contribution of each variable to the MIR. Variables with more influence over the MIR are determined through analysis of the beta scores. The significant contributors to the MIR are determined first by identifying the variables with p-values less than 0.05. The thirteen significant indicators are listed below in Table 5.1 in order from highest to lowest beta value.

Further analysis involves looking at the contribution of the significant indicators within the conceptual model. The role of each indicator in the vulnerability of populations is important and has implications to the results obtained in later grouping methods. The significant variables that showed up in the model did appear to align with each of the four factors from the proposed model from Figure 2.3 and will be discussed in more detail Chapter 6.
Table 5.3. Indicators with significant correlations to the MIR in the 34 variable regression model. Listed in order from highest to lowest beta values.

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Beta Value</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obesity</td>
<td>0.176</td>
<td>0.000</td>
</tr>
<tr>
<td>Rural</td>
<td>0.115</td>
<td>0.000</td>
</tr>
<tr>
<td>Single parent households</td>
<td>0.097</td>
<td>0.001</td>
</tr>
<tr>
<td>Low birth weight</td>
<td>0.094</td>
<td>0.000</td>
</tr>
<tr>
<td>Unmarried</td>
<td>0.095</td>
<td>0.002</td>
</tr>
<tr>
<td>Education level</td>
<td>0.081</td>
<td>0.028</td>
</tr>
<tr>
<td>Smokers</td>
<td>0.079</td>
<td>0.000</td>
</tr>
<tr>
<td>No religious affiliation</td>
<td>0.075</td>
<td>0.000</td>
</tr>
<tr>
<td>Lack of exercise</td>
<td>0.068</td>
<td>0.017</td>
</tr>
<tr>
<td>No mammogram in 2 years</td>
<td>0.053</td>
<td>0.010</td>
</tr>
<tr>
<td>Lack of Health Food Access</td>
<td>0.049</td>
<td>0.003</td>
</tr>
<tr>
<td>Income Inequality</td>
<td>0.049</td>
<td>0.039</td>
</tr>
<tr>
<td>Lack of Binge Drinking</td>
<td>0.041</td>
<td>0.040</td>
</tr>
</tbody>
</table>

The regression analysis indicates that the most influential variables on the MIR mainly stem from the general health/behavior and social characteristics. Smoking, drinking, obesity, and lack of exercise all have high and significant beta values that suggest greater contributions to cancer outcomes related to the health of an individual as well as their behaviors or habits. Of these, obesity is the most compelling and reflects much of what is known about the connection between health and being obese. Single parent households, religious affiliation and the percentage of unmarried individuals in a county also show up as more influential variables in the data set. These variables all suggest the importance of social support and its potential impact on treatment success. Percentage of a county designated as rural also has a higher beta value and may be a measure of support, in addition to acting as a proxy for access. The lowest significant beta values in the set come from health food access and, income inequality and the binge-drinking variables. Education level also possesses a higher beta value in addition to a
higher VIF score. Although this variable does not possess the highest beta value with relation to the MIR, it does correlate highly with a number of the variables in the set, including a few of the more influential and making it a potentially strong indicator when looking at underlying causes.

The goal of the first regression model was to assess the relationship of individual indicators to the MIR in addition to determining how the indicators interact with each other within the conceptual model to produce higher or lower vulnerability to cancer fatality. This establishes a baseline to use in determination of the adequacy of the grouping methods and the conceptual model. The adjusted $R^2$ suggests a far more complicated relationship between the indicators and the MIR, and the objective is to determine which variables, or groups of variables, most influence the complexities. It is likely that the same variables influencing the predictive value of the set are the same variables responsible for the disparities. Path and spatial analyses help to discern some of the relationships between the grouped variables in the next two chapters.

One of the only real surprises in the analysis of individual variables was the lack of influence from race. The percent non-white variable was expected to demonstrate at least a modest influence over the MIR based on a number of studies showing a link between race and cancer outcomes. The correlation between the non-white variable and the MIR was only 0.296, however. This would indicate a weak relationship at best. Considering both the cancer survivorship among African American populations, which is relatively low when compared to the white population, and the types of cancer, which are most likely in the population, it seems as if there should be a more robust relationship
present. The results in this research do not support many of the findings, however, that race or ethnicity plays a major role in cancer outcomes.

Each indicator was adjusted prior to entry to ensure a positive relationship with the MIR, i.e. as the indicator value goes up, so does the MIR. There are ten indicators, however, that have negatively correlate with the MIR, including two that are significant. The indicators are non-English speaking, renters, non-insured, public parks per capita (inverse), recreation centers per capita (inverse), liquor stores per capita, health food access (inverse), TRI output per capita, and binge drinking. It seems like each of these variables has an association with urban environments. Looking at the correlation matrix, however, this relationship cannot be consistently confirmed. Only a few of these variables are negatively correlated with the rural indicator at any substantial level. These are adjusted for in the grouping of variables in the next two chapters so that all indicators have a positive correlation to the MIR.
CHAPTER 6

PREDOMINANT BROAD-BASED VULNERABILITY FACTORS

This chapter focuses on the second research question, involving identification of the broad-based factors that account for most of the spatial variability in cancer outcomes. The broad-based factors are intended to represent real-life paradigms and are useful in the operationalizing of the disparity model. Variables were grouped through both the use of a priori knowledge to match the model parameters and through statistical methods (inductive grouping). Testing of the groupings was accomplished through both a regression analysis and a comparison of the two models, theoretically grouped and inductive grouped, to each other.

6.1. Factor Creation through Theoretical Grouping of Variables

The theoretical grouping of variables accounted for established societal structures and the expectations of how these variables would play out in relation to each other. There are four sets of characteristics that form the major themes of influence on cancer vulnerability, including: social characteristics, general health and behavioral characteristics, financial and medical access characteristics, and community and environment characteristics. The thirty-four original variables were grouped into these four categories, shown below in Table 6.1a and 6.1b.
Table 6.1a. 34 variables separated into 4 theoretically grouped factors. Pearson’s R shown for each individual variable compared to the group value along with correlations between the groups.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Variable</th>
<th>R</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social Characteristics</td>
<td>Lower Religious Affiliation</td>
<td>0.234</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Reporting no social support</td>
<td>0.601</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Unmarried</td>
<td>0.778</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Non-white</td>
<td>0.791</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Single Parent Households</td>
<td>0.799</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Language Isolation</td>
<td>0.186</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Renters</td>
<td>-0.426</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Households with dependents</td>
<td>0.384</td>
<td>0.000</td>
</tr>
<tr>
<td>Health and Behavioral</td>
<td>Obesity</td>
<td>0.716</td>
<td>0.000</td>
</tr>
<tr>
<td>Characteristics</td>
<td>Reporting health as “poor”</td>
<td>0.519</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Low Birth Weight</td>
<td>0.610</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Lack of exercise</td>
<td>0.804</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Percent High Risk Occupations</td>
<td>0.324</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Smokers</td>
<td>0.597</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>No mammogram in last 2 years</td>
<td>0.555</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Alcohol consumption (binge)</td>
<td>0.472</td>
<td>0.000</td>
</tr>
<tr>
<td>Financial and Medical Access</td>
<td>Income Inequality (GINI)</td>
<td>0.282</td>
<td>0.000</td>
</tr>
<tr>
<td>Characteristics</td>
<td>Unemployment Rate</td>
<td>0.575</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Mammogram Facilities</td>
<td>0.472</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Median household income</td>
<td>0.546</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td># of doctors</td>
<td>0.649</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Oncology facilities</td>
<td>0.449</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Uninsured population</td>
<td>-0.184</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td># of internal medicine doctors</td>
<td>0.498</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Educational attainment</td>
<td>0.527</td>
<td>0.000</td>
</tr>
<tr>
<td>Community and Environmental</td>
<td>Number of Public Parks</td>
<td>0.509</td>
<td>0.000</td>
</tr>
<tr>
<td>Characteristics</td>
<td>Recreational Facilities</td>
<td>0.360</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Rural/Urban mix</td>
<td>0.482</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Population Growth</td>
<td>0.009</td>
<td>0.630</td>
</tr>
<tr>
<td></td>
<td>Population Density</td>
<td>-0.041</td>
<td>0.029</td>
</tr>
<tr>
<td></td>
<td>Liquor Store Density</td>
<td>0.385</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Fast Food Density</td>
<td>-0.162</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Healthy food access</td>
<td>0.371</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Environmental Exposure (TRI)</td>
<td>0.030</td>
<td>0.108</td>
</tr>
</tbody>
</table>
Table 6.1b. Pearson’s R correlations between each of the 4 theoretically grouped factors.

<table>
<thead>
<tr>
<th></th>
<th>Social</th>
<th>Health and Behavioral</th>
<th>Financial and Medical Access</th>
<th>Community and Environmental</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social</td>
<td>1.00***</td>
<td>0.37***</td>
<td>0.29***</td>
<td>0.27***</td>
</tr>
<tr>
<td>Health and</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Behavioral</td>
<td>1.00</td>
<td>0.59***</td>
<td>0.01***</td>
<td></td>
</tr>
<tr>
<td>Financial and</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medical Access</td>
<td>1.00</td>
<td></td>
<td>-0.01***</td>
<td></td>
</tr>
<tr>
<td>Community and</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Environmental</td>
<td></td>
<td></td>
<td></td>
<td>1.00</td>
</tr>
</tbody>
</table>

***Significant at p.001 level

The correlations in table 6.1a and 6.1b provide some initial insight into the validity of the groupings. Higher correlations are indicative of a better fit for the variable within the broad-based factor. The general health variables appear, from this analysis, to be the most highly correlated. There are six significant correlations of greater than 0.5 and no negative correlations in this factor. Social variables also have a higher number of significant correlations of greater than 0.5, however, the renter variable has a negative correlation, meaning it likely measures the homeownership as opposed to renters. The financial and medical access factor does not have correlation values over 0.7 like the previous factors, but still does have four with correlations of greater than 0.5. Only the uninsured population in this factor has a negative correlation, indicating that the variable should measure the insured population in order to align with the other variables. The final factor, intended to represent the community and environmental characteristics of a place, does not appear to be well constructed based on the data in table 6.1a and b. There is only one variable with a correlation of greater than 0.5. In addition, there are two negative correlations in the group and two statistically insignificant correlations.
Following the correlation analysis, a regression model was run using the same method as with the entire variable set. To test the predictive ability of the four-factor set, the adjusted $R^2$ was analyzed and compared to the original 34-variable model. In addition, a Durbin-Watson statistic was calculated to determine the level of autocorrelation appearing within the set. The results of the regression model are presented in Table 6.2, while the betas values, significance levels and VIFs are in Table 6.3.

Table 6.2. Theoretically Grouped Factors Regression model results.

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Standard Error of the Estimate</th>
<th>Durbin-Watson</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.570a</td>
<td>.325</td>
<td>.324</td>
<td>.8057218</td>
<td>2.017</td>
</tr>
</tbody>
</table>

Table 6.3. Theoretically Grouped Factors Explanatory variables

<table>
<thead>
<tr>
<th>Model</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
<th>Collinearity Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Constant)</td>
<td></td>
<td>- .422</td>
<td>.673</td>
<td></td>
</tr>
<tr>
<td>Social Characteristics</td>
<td>.190</td>
<td>10.144</td>
<td>.000</td>
<td>.795</td>
</tr>
<tr>
<td>Health and Behavioral</td>
<td>.410</td>
<td>19.542</td>
<td>.000</td>
<td>.817</td>
</tr>
<tr>
<td>Medical and Financial Access</td>
<td>.092</td>
<td>6.557</td>
<td>.000</td>
<td>.795</td>
</tr>
<tr>
<td>Community and Environmental</td>
<td>.073</td>
<td>4.430</td>
<td>.000</td>
<td>.995</td>
</tr>
</tbody>
</table>

Adjusted $R^2$ values, f-statistics and VIF values of the theoretically grouped variable model were compared to the original 34-variable model to determine the effectiveness of grouping the variables. The adjusted $R^2$ of the model was 0.324, which did not represent a major difference from the 34-variable model, with an adjusted $R^2$ of
0.347. The Durbin-Watson statistic also indicates little autocorrelation in the model. In addition, the VIF values are all very close to one, signifying little multicollinearity. Each of these three model indicators provides evidence that the groupings are well constructed.

The final test of grouping adequacy was conducted by running a Cronbach’s alpha analysis. This test was intended to confirm the internal consistency of each factor by creating sets of correlation pairings within the group and follow up on the findings of the original correlation analysis. The resulting alpha statistic ranges from 0, indicating no consistency within the group, to 1, indicating perfect consistency. A value of greater than 0.6 is considered significant, and indicates the potential that stronger relationships exist among the variables in the group (Cronbach 1951). In other words, the variables represent the broader construct. Results of the Cronbach’s tests are in Table 6.4.

Table 6.4. Cronbach’s Alpha for theoretically grouped factors.

<table>
<thead>
<tr>
<th>Group</th>
<th>Cronbach’s Alpha</th>
<th>N of items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social Characteristics</td>
<td>.601</td>
<td>7</td>
</tr>
<tr>
<td>Health and Behavioral</td>
<td>.699</td>
<td>9</td>
</tr>
<tr>
<td>Medical and Financial Access</td>
<td>.462</td>
<td>9</td>
</tr>
<tr>
<td>Community and Environmental</td>
<td>-.071</td>
<td>9</td>
</tr>
</tbody>
</table>

Two of the factors, social characteristics and health and behavioral characteristics, had significant alpha values. This suggests that these groups are composed of variables that are strongly correlated to each other. Medical and financial access, while not significant, still had an alpha value indicative of some correlation amongst the variables in the factor. The alpha value for the community and environmental factor, which was exceptionally low, suggested that at least some of the variables in the factor have a
negative correlation with each other and that some of the a priori assumptions about the factor are incorrect.

In addition to testing the adequacy of the factor groupings, the regression model run using the broad-based factors was used to test the strength of the relationship between each factor and the MIR. Instead of single variables, the goal was to find how different components of a place influence cancer fatality. Identification of patterns within the data provided insight into how the characteristics function in society. When trying to reduce cancer fatalities or eliminate disparities, knowledge of these patterns will be critical.

Another important aspect of the broad based factor analysis was the description of each factor. In other words, did any of the variables within each factor appear to exert more influence or provide a connection between other variables within the factor? This was conducted through analysis of the most significant contributors to the MIR from each factor, correlations existing within each factor, and the relationship of each factor to the MIR.

The health and behavioral characteristics had the highest correlation with the MIR based on the beta value from the regression model. (Table 6.3) Within this group, obesity and exercise were the most highly correlated, and also happened to have significant correlations to the MIR (shown in regression results from Table 5.1). Social characteristics had the next highest correlation to the MIR as well as the second highest level of internal consistency. Within this factor, the single parent household variable has the highest individual correlation to the MIR along with the highest correlation to the factor score. Unmarried populations also correlate highly, both with the MIR and the factor score. Medical and financial access characteristics are represented primarily by the
education level and number of doctors. Education level had a significant correlation with the MIR in the 34-variable regression and the number of doctors correlates most highly within the factor. The doctor variable also correlates highly (R=0.74) with the education variable. The community and environmental characteristics factor had the lowest correlation to the MIR, but also had the lowest measure of internal consistency (Table 6.4). Within this factor, the percentage of a county designated as rural had the highest correlation to the MIR, and to the community factor.

6.2 Factor Creation through Inductive Grouping of Variables

After grouping the variables theoretically and testing them in the regression model, the thirty-four individual variables were input into a PCA model. The benefit of the PCA analysis was that the dimensionality of the data set would be reduced in a manner that created orthogonal factors that explained a majority of the variability within the data set. The goal of the dimension reduction was to both identify sets of characteristics that exerted the most influence on disparities and to create a set of factors to which the theoretical groupings could be compared. Characteristics that both have the most variation and the highest correlation to the MIR should have the most significant contribution to disparities. In addition to the above-described reasons for using the inductive method, another benefit of the PCA was in the identification of variable groupings not previously established by research.

In order to assure that the resulting factors in the PCA were orthogonal, a varimax rotation was used for factor extraction. Two different factor retention approaches were then assessed prior to a decision being made on factor retention. The first approach
involved retention of factors based on Kaiser Criterion, meaning they had an eigenvalue of greater than one. This resulted in the retention of ten factors explaining 63.43% of the variance within the data set. The second approach involved the theoretical extraction of factors based on scree plot analysis, looking for a change in slope to determine the number of factors to extract. The slope from factors one to five was relatively consistent with a slight decrease in slope from five to six. Beyond factor six, the scree plot leveled out, indicating that six factors should be retained. The resulting six-factor solution explained only 46.56% of the variance within the data set, however.

Loading plots for the rotated factor solutions in each of the PCA models were then evaluated to determine the number of variables having a significant load on each factor, a value of greater than 0.5. In order for a factor to be retained, there needed to be at least two variables loading significantly on that factor. Using Kaiser Criterion, eight of the ten factors had at least two variables with significant loadings. The scree plot method only produced four factors with any significant variable loading. Of the four, only the first two had more than one variable with a significant factor loading. As a result of both the significant variable loadings and the higher level of variance explained, the Kaiser Criterion factor retention model was retained and ten factors were created.

Following the choice of factor extraction method, the next step in the inductive method was to determine societal constructs represented by each factor in the model. The rotated component matrix was looked at and each variable with a significant loading was highlighted. Table 6.5 below shows the ten principle components along with the variables loading significantly on each. A brief description of the societal construct most likely described by each factor is also included in this analysis. There is also an attempt to
describe each of the identified PCA factors in relation to the theoretically grouped factors.

Table 6.5. Results of PCA run on 34 variables. Factors are created using varimax rotation and retained using Kaiser Criterion (eigenvalue > 1). Variance contribution of each group is displayed along with the correlation coefficient of each variable from the rotated component matrix. Factors listed are those with values > 0.5.

<table>
<thead>
<tr>
<th>Factor (explained variance)</th>
<th>Variable</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factor 1 – 18.4%</td>
<td>Doctors</td>
<td>0.796</td>
</tr>
<tr>
<td></td>
<td>Oncologists</td>
<td>0.776</td>
</tr>
<tr>
<td></td>
<td>Education Level</td>
<td>0.701</td>
</tr>
<tr>
<td></td>
<td>Rural</td>
<td>0.692</td>
</tr>
<tr>
<td></td>
<td>Homeowners</td>
<td>0.655</td>
</tr>
<tr>
<td></td>
<td>Fast Food Access</td>
<td>-0.670</td>
</tr>
<tr>
<td>Factor 2 – 12.4%</td>
<td>Non-white population</td>
<td>0.885</td>
</tr>
<tr>
<td></td>
<td>Single parent households</td>
<td>0.832</td>
</tr>
<tr>
<td></td>
<td>Unmarried</td>
<td>0.756</td>
</tr>
<tr>
<td></td>
<td>Low Birth Weight Babies</td>
<td>0.691</td>
</tr>
<tr>
<td></td>
<td>Obese population</td>
<td>0.500</td>
</tr>
<tr>
<td>Factor 3 – 6.5%</td>
<td>Dependents</td>
<td>0.650</td>
</tr>
<tr>
<td></td>
<td>Income Inequality</td>
<td>0.587</td>
</tr>
<tr>
<td></td>
<td>Household Income</td>
<td>0.577</td>
</tr>
<tr>
<td></td>
<td>Regular mammogram</td>
<td>0.573</td>
</tr>
<tr>
<td>Factor 4 – 5.5%</td>
<td>Poor Health</td>
<td>0.740</td>
</tr>
<tr>
<td></td>
<td>Smokers</td>
<td>0.725</td>
</tr>
<tr>
<td></td>
<td>No exercise</td>
<td>0.509</td>
</tr>
<tr>
<td>Factor 5 – 4.2%</td>
<td>Oncology units</td>
<td>0.734</td>
</tr>
<tr>
<td></td>
<td>Mammogram facilities</td>
<td>0.726</td>
</tr>
<tr>
<td>Factor 6 – 3.8%</td>
<td>Liquor Stores</td>
<td>0.669</td>
</tr>
<tr>
<td></td>
<td>Number of Parks</td>
<td>0.566</td>
</tr>
<tr>
<td>Factor 7 – 3.3%</td>
<td>Religious Adherence</td>
<td>0.773</td>
</tr>
<tr>
<td></td>
<td>High Risk occupations</td>
<td>0.512</td>
</tr>
<tr>
<td>Factor 8 – 3.2%</td>
<td>TRI Releases</td>
<td>0.663</td>
</tr>
<tr>
<td>Factor 9 – 3.1%</td>
<td>Population Growth</td>
<td>0.670</td>
</tr>
<tr>
<td></td>
<td>Density</td>
<td>-0.561</td>
</tr>
<tr>
<td>Factor 10 – 3.0%</td>
<td>Health Food Access</td>
<td>0.791</td>
</tr>
</tbody>
</table>

Paired t-tests (Tables 6.6- Table 6.8) are used to compare the factors statistically along with analysis of the correlations within each factor, shown in Table 6.1.
Conducting this comparison of the factor compositions was a good way to account for any discrepancies within the theoretically grouped factors and to help in the formation of a more robust model for cancer disparities.

Factor one, accounting for over 18% of the variance within the set, contains the significant variables measuring extent of access to medical care within the population. The variables within this group are recognizable from the medical and financial access characteristics as well as the community and environment characteristics in the theoretically grouped factors. To test the relationship of the two factors, a paired t-test was conducted and results are shown in Table 6.6. The results demonstrate that the two factors are statistically similar. Based on this information, the most likely representation for factor 1 is the access to health care providers.

Table 6.6. Paired t-test for access characteristics factor and PCA factor 1. (α = 0.05)

<table>
<thead>
<tr>
<th>Paired Differences</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Std. Error Mean</th>
<th>95% Confidence Interval of the Difference</th>
<th>t</th>
<th>df</th>
<th>Sig. (2-tail)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pair 1</td>
<td>.0234</td>
<td>1.1889</td>
<td>.02276</td>
<td>-.02114 to .06811</td>
<td>1.032</td>
<td>2727</td>
<td>.302</td>
</tr>
</tbody>
</table>

Factor two, accounting for over 12% of the variance in the set, contains significant indicators measuring both the family structure (single parents and unmarried), and health indicators (obesity and low birth weight). The relationship of this PCA factor is tested in comparison to both the social characteristics factor and the health and behavioral characteristics factor from the theoretically grouped set due to the almost even
split of the variables within the factor. The result of the two paired t-tests, shown below in table 6.7, provides evidence that the construct represented by factor two is related more to the general health of the population.

Table 6.7. Paired t-test for health and social characteristics factors with PCA factor 2. ($\alpha = 0.05$)

<table>
<thead>
<tr>
<th>Paired Samples Test</th>
<th>Paired Differences</th>
<th>95% Confidence Interval of the Difference</th>
<th>t</th>
<th>df</th>
<th>Sig. (2-tail)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Std. Dev.</td>
<td>Std. Error Mean</td>
<td>Lower</td>
<td>Upper</td>
</tr>
<tr>
<td>Pair 1 Social PCA-Fac. 2</td>
<td>.02348</td>
<td>.64981</td>
<td>.01244</td>
<td>-.00090</td>
<td>.04788</td>
</tr>
<tr>
<td>Pair 2 Health PCA-Fac. 2</td>
<td>.00071</td>
<td>.92248</td>
<td>.01766</td>
<td>-.03391</td>
<td>.03534</td>
</tr>
</tbody>
</table>

Factor 3 accounts for 6.5% of the variance in the data set. Unlike the previous two factors, however, there is a combination of three theoretically grouped factors represented. Social characteristics are represented by the dependents variable. Medical and financial access characteristics are represented by median household income and income inequality. Health and behavioral characteristics are represented by the regular mammograms variable. Based on a simple visual analysis, the factor seems to be represented by the family structure, with number of dependents having the highest correlation in the component matrix. A paired t-test, shown below in Table 6.8, reveals a stronger relationship to the health characteristics than the other measures. This indicates that factor 3 is likely measuring likelihood of regular medical care and screening.
Table 6.8. Paired t-test for social, access and health and characteristics factors with PCA factor 3. (α = 0.05)

<table>
<thead>
<tr>
<th>Paired Differentials</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Std. Error Mean</th>
<th>95% Confidence Interval of the Difference</th>
<th>t</th>
<th>df</th>
<th>Sig. (2-tail)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pair 1 Social PCA-Fac. 3</td>
<td>.02348</td>
<td>.99581</td>
<td>.01906</td>
<td>-.01389</td>
<td>.06087</td>
<td>1.232</td>
<td>2,727</td>
</tr>
<tr>
<td>Pair 2 Access PCA-Fac. 3</td>
<td>.00481</td>
<td>.97126</td>
<td>.01859</td>
<td>-.03165</td>
<td>.04127</td>
<td>.259</td>
<td>2,727</td>
</tr>
<tr>
<td>Pair 3 Health PCA-Fac. 3</td>
<td>.00071</td>
<td>.92899</td>
<td>.01778</td>
<td>-.03416</td>
<td>.03559</td>
<td>.040</td>
<td>2,727</td>
</tr>
</tbody>
</table>

Factors four and five align very well with the theoretically constructed factors. Factor 4, which accounts for 5.5% of the variance in the data set, is composed of smoking, poor health, and no exercise. Each of these variables comes from the health and behavioral characteristics factor, but these three in particular seem related to negative health behaviors. Factor five accounts for 4.2% of the variance and contains the variables measuring the number of oncology and mammography units in a county. Based on the construct measured by these, factor five could be considered a measurement of facilities access.

Factor six, accounting for 3.8% of the variance in the set, contains two variables from the theoretically constructed community and environment factor. Liquor stores and number of parks constitute the two significant variables in this factor, and correlate positively with each other, meaning counties with more parks also have more liquor.
stores. The only possible explanation for the statistical relationship between these two is a potential tie to urban areas. Placing these two into the same factor, however, does not provide much explanatory power with respect to the MIR. Table 6.9 below, shows the low beta value and lack statistical significance for this factor in the regression model for the PCA factors.

The last four factors will be discussed together based on their significance in the regression model (Table 6.9). Of the four, factors seven and nine are not significantly correlated to the MIR in the regression model. Religious adherents and high-risk occupations represent factor seven, while factor nine is represented by population growth and population density. Factors eight and ten both have significant relationships to the MIR. Factor eight is represented by only the TRI variable and factor ten is represented by only healthy food access, which makes these two easily identifiable.

Following the creation of ten factors created using the inductive method, another regression model was run to determine the predictive ability of the entire set on the MIR as well as the correlation of each factor to the MIR. The factors scores used in the model are constructed using a summation the values multiplied by the loading weights of each variable within the factors. Results of the model, shown below in table 6.9, yielded an adjusted $R^2$ of 0.332 and a Durbin-Watson statistic of 2.008. This indicates that much of the explanatory power of the set is retained in addition to having very little autocorrelation amongst the factors. Also, as expected, the VIF values for each factor are one. This is the result of using a varimax rotation in the factor extraction process. Each of the factors in the model is represented in this table by the most likely model construct.
embodied. This helps to add some context to the analysis of significant relationships to
the MIR.

Table 6.9. Regression results for PCA-grouped variables using calculated factor scores.
**Significance at p < 0.01, *Significance at p < 0.05

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
<th>Collinearity Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td></td>
<td>Tolerance</td>
</tr>
<tr>
<td>1-Access to Health Care</td>
<td>.255</td>
<td>.015</td>
<td>.261**</td>
<td>16.659</td>
<td>.000</td>
</tr>
<tr>
<td>2-Gen. Health Characteristics</td>
<td>.352</td>
<td>.015</td>
<td>.359**</td>
<td>22.951</td>
<td>.000</td>
</tr>
<tr>
<td>3- Regular Care / Screening</td>
<td>.174</td>
<td>.015</td>
<td>.178**</td>
<td>11.362</td>
<td>.000</td>
</tr>
<tr>
<td>4-Negative Health Behavior</td>
<td>.280</td>
<td>.015</td>
<td>.286**</td>
<td>18.245</td>
<td>.000</td>
</tr>
<tr>
<td>5-Access to Medical Facilities</td>
<td>.038</td>
<td>.015</td>
<td>.039</td>
<td>2.503</td>
<td>.012</td>
</tr>
<tr>
<td>6-Urban areas</td>
<td>.030</td>
<td>.015</td>
<td>.030*</td>
<td>1.938</td>
<td>.053</td>
</tr>
<tr>
<td>7-Religious adherence</td>
<td>.004</td>
<td>.015</td>
<td>.004</td>
<td>.241</td>
<td>.810</td>
</tr>
<tr>
<td>8-Environment Hazards</td>
<td>.131</td>
<td>.015</td>
<td>.134**</td>
<td>8.542</td>
<td>.000</td>
</tr>
<tr>
<td>9-Population Growth</td>
<td>.010</td>
<td>.015</td>
<td>.011</td>
<td>.681</td>
<td>.496</td>
</tr>
<tr>
<td>10-Health Food Access</td>
<td>.062</td>
<td>.015</td>
<td>.064**</td>
<td>4.065</td>
<td>.000</td>
</tr>
</tbody>
</table>

The regression model created by PCA grouped variables did mark some
improvement over the theoretically grouped model with regards to the overall predictive
power. Despite the differences in how variables grouped in each of the methods, some of
the same patterns emerged in relation to the MIR. Similarly to the theoretically grouped factor regression model, the health and behavioral variables had the highest beta values in the model. Health characteristics such as obesity and low birth weight, as well as behavioral characteristics like smoking and lack of exercise, again exhibit a strong correlation to the MIR. The highest beta value in the model came from factor two, which is representative of general health characteristics.

An interesting divergence between the grouping methods came from the medical and financial access factor. In the theoretically grouped regression model, the second highest beta value came from the social characteristics. In the inductive model, however, the access factor appears to be more highly correlated to the MIR. There are social characteristics present in the first three factors of the PCA. In fact, these variables represent a majority of the correlation to the MIR. This influence, shown below in Figure 6.1, contributes to the higher beta values in the regression model. Non-white population, single parent households, unmarried population and dependents are all accounted for in the model, yet are split up amongst the health and access characteristics. This could be indicative of an underlying relationship between the social fabric of a place and the resulting health and access characteristics present. A path analysis, which is discussed in section 6.4, helps to flesh out some of the possible relationships leading to cancer disparities.

Unlike the analysis of theoretically grouped indicators using a priori knowledge of societal constructs and influence on cancer, the inductive analysis provided a different perspective on the relationship between the variables. The groupings, when analyzed, displayed combinations of variables from multiple a priori constructs. For example,
factor one was represented by a combination of both access characteristics (doctors and education level) and community characteristics (rural and fast food). The PCA revealed that statistical grouping of variables did not follow the same conceptual groupings, putting variables from the three constructs together in some cases. This result uncovers a potential link between factors that influences cancer disparities. Going back to the proposed model, the health outcome in a place results from a combination of all factors present in that place. Certain factors may correlate highly with other factors due to a path of influence. In other words, the presence of one characteristic may increase the likelihood of another set of characteristics being present. For example, a higher education level (access factor) may lead to higher income levels (also access), which will lead to a higher likelihood of access to recreational facilities and healthy food. All of this may result in a higher chance of exercising and eating well and a lower vulnerability to negative cancer outcomes. To summarize, even though the model constructs (factors) may be constructed accurately, the relationship between the factors may be directionally dependent, meaning that one or more factors are ultimately responsible for starting a chain of events that leads to higher cancer fatality rates. The possibility of chains of events influencing cancer outcomes is tested through a path analysis, discussed in section 6.4.

6.3. Summary of Group Findings

The broad based factor question is answered through the principle components, grouped regression, and reliability analyses. Factors are represented through theoretically constructed groups as well as PCA-constructed groups to establish patterns of significant
interactions among the indicators. The results of both the theoretically grouped and the PCA grouped variables revealed little reduction in predictive power of the model, based on the adjusted $R^2$, when compared to the regression with thirty-four individual variables. Between both approaches, there was also evidence of some consistency within the groups. Figure 6.1 displays the relative influence of each theoretically grouped factor along with beta scores and the individual variables with significant correlations to the MIR. The size of each oval in the figure is proportionate to the correlations with the MIR. Figure 6.2 displays the relative contribution of each indicator to the variance determined from the PCA model. The proportional ovals are scaled to represent the contribution of each factor to the MIR as well as the contribution of significant loading variables to the factor. The figures help to visualize the contributions of each group and significant indicator along with the potential sources of disparities. Following the diagrams is a summary of the research findings within each of the four theoretically constructed groups, taking into consideration both the regression and PCA results.
6.3.1. Health and Behavioral Characteristics

The group representing health and behavioral characteristics of a place was found to have a majority of the influence over MIR within the theoretically grouped regression model. The variables in this factor also tended to group together in the PCA factors, loading together heavily on factors two, three, four and ten. In addition, this group had the highest internal consistency among the individual variables and contained the most significant independent variables from the 34-variable regression model.
Figure 6.2. Diagram depicting the 10 PCA constructed factors. Also shown are the significant variables contributing to each factor, if there are any. Ovals are all proportionate to their beta values in the regression model. Variables are color coded to denote the theoretical groups to which each variable belongs. Note: diagram not drawn to exact scale.

Understanding the composition of the factor is important to understanding the link between amongst the variables and to the MIR. Exploration of the variables within this factor revealed a few major contributors to the MIR, including obesity and low birth weight babies. Obesity has well-established links to cancer and many other health problems, so its presence came as little surprise. The low birth weight variable was included in the analysis because of the known association with adverse health outcomes. While some studies link low birth weight to cancer and obesity later in life, the inclusion
of this variable was meant as a possible proxy for the health of the mothers. The reasons for low birth weight babies typically stem from maternal health issues such as poor nutrition, smoking, high stress, and diseases or infections. All of these issues are also risk factors for cancer, making the low birth weight a relatively strong predictor variable for the health of the female population. This particular indicator may prove to be a very strong predictor in female cancers such as breast, ovarian, or cervical.

Rounding out the top indicators in this factor were smokers and women over forty not having a mammogram in the last year. Smoking is surprising in that it didn’t have more influence considering the strong ties between cigarettes and cancer. It was beat out by a considerable margin by both obesity and low birth weight, however. Mammograms also have a relatively strong correlation to MIR. It would have been helpful to include the prostate specific antigen (PSA) test as well, but unfortunately the data for this test is not collected and available at the county level as it for the mammogram variable. Along with breast cancer, prostate cancer makes up one of the most readily diagnosed cancers and is also responsible for killing almost 150,000 men annually. The all-cancer MIR and the proxies are likely influenced heavily by the presence of these two cancers. Future models including only breast and prostate cancer would likely substantiate similar correlations, and potentially produce stronger model fits.

The importance of the findings from this group is the confirmation of influence coming from the obesity and low birth weight indicators in this group. Both of these indicators displayed both a high correlation to the MIR in the regression analysis as well as a significant contribution to the variability in the data set, denoting their importance in disparities.
6.3.2. Social Characteristics

Social characteristics represent the relationships between individuals and the community as a whole, and had the second highest influence over the MIR in the group regression model. The variables in this group represent characteristics that may influence an individual’s ability to cope with problems. Analysis of the correlations between the individual variables in this group and the MIR, by far the most significant predictor is the single parent household. The other two significant correlations come from unmarried populations and religious adherents. Each of these variables seems to measure a similar concept with regards to cancer care. The individuals represented by these statistics may be less likely to get screening tests, and if diagnosed, less likely to go through extended treatment. Social networks and support systems have proven health benefits are crucial when dealing with a chronic illness. The correlation between these indicators and negative cancer outcomes is not surprising.

The variance in the social characteristics is relatively large, as is evidenced by the number of variables in this set possessing a significant loading on the first three factors identified in the PCA model. Single parent households, renters, non-married and non-white populations are the major contributors. Of these, the single parent household is the only one also with a significant correlation to the MIR, making it important in both its connection with cancer deaths and with disparities.

6.3.3. Financial and Medical Access

Financial access characteristics had a slightly lower beta value than the social characteristics. Also, there were only two significant variables in the factor, education
level and income inequality. Surprisingly, the income level indicator had the lowest beta value in the factor along with insurance. Further complicating findings, education level has a strong correlation to income level. This indicates that education does not lower cancer vulnerability by increasing income levels, and may not be a factor in access. Instead, the relationship may be filtered through either the social or health indicators. To confirm these possibilities, correlations are analyzed between the education level variable and the individual social and health variables. It is evident from this examination that education does have a connection to health behaviors, with a negative correlation to obesity, exercise, and smoking. There is no correlation to any of the social indicators.

The number of doctors and oncologists, as well as the number of facilities located in a county did not have a significant influence. This was not unexpected, as it does not matter how many doctors and hospitals are surrounding an area if the population lacks the ability to pay for the care.

The relevant connection between financial access and cancer fatality is likely related to knowledge. Higher education levels may result in more awareness of cancer risks and lead to the less engagement in high-risk behaviors and more engagement in low-risk behaviors. This could, in turn, lead to later diagnosis of cancer or to more aggressive cancers.

6.3.4. Community and Environment

This group proved to be the least predictive and have the least internal inconsistency. Analysis of the individual indicators within the group reveals two with significant correlations to the MIR. Percentage of rural area in a county has the highest of the beta values in this group, and actually has the second highest of all individual
indicators. Health food access is the other significant indicator, but likely measures a similar construct to that of rural areas.

There were a number of unexpected relationships evident in the analysis as well. The variables for parks, recreational facilities, and health food all positively correlate with the MIR. This means, if taken out of the context of the research, more parks, recreational facilities, and health food stores are related to higher cancer fatality rates. This, of course, does not make sense when considering the other variables. The more probable explanation is that rural areas have less of these features in addition to having higher cancer fatality. Analysis of the correlations within the group confirms this, with rural areas negatively correlated to each of these variables, although weakly. Rural areas have less medical facilities and doctors, while also having higher correlations with unhealthy behaviors such as lack of exercise and smoking. In addition, there is a relatively strong negative correlation to education level. Considering all of this, the community group is probably defined by access and measured by rural/urban proxy.

Based on the findings from both the theoretically grouped and PCA grouped factor analyses, there is not enough evidence to support changing the grouping of variables. Using a priori methods to group the data retains most of the predictive power of the original data set, as evidenced by the adjusted $R^2$ values from the regression models. The PCA grouping method sheds some light on potential interactions between the factors, demonstrating some different arrangements of variables when compared to the theoretical groupings. The only possible change merited by the results would be to rearrange the community variables and either dissolve the group into the others, or
remove some of the variables with little explanatory power and negative correlations to other variables in the factor. This may be a worthwhile venture for future research.

6.4. Verification of Group Influence on MIR – Path Analysis

Modifiable health risks, or decisions made that impact the health of individuals, are generally composed of a series of events that are connected to each other. For example, living in lower socioeconomic area may result from having less education and lead to less health food access. A path representing this example would begin with the education variable, lead through income level, ending with health food access. The ultimate source of the vulnerability is an important part to identify in this research. The source identifies where the chain of events leading to cancer death begins, and where the resources should be focused in order to break the chain and improve the health of the community. Filtering through certain characteristics may increase or decrease the impact of other characteristics.

Path analysis allowed for the investigation of multiple chains of events that potentially lead to cancer fatality. The goal was to test the validity of the conceptual model and to identify significant paths to cancer fatality. There is a distinction between how characteristics of a place are grouped into societal constructs and how they are grouped to measure cancer outcomes. This was evident in the difference between the two grouped regression models. The path analysis maintained the societal constructs of the theoretically constructed model while also allowing the interaction of variables evident in the PCA grouped model to be tested. The resulting path model is shown in Figure 6.3. The four groups in the figure are composed of the original thirty-four variables, and the numbers between the groups represent path coefficients and correlations between them.
The adjusted $R^2$ for the analysis is displayed above the MIR variable and represents the predictive ability of the model. Each of the four groups represents a correlated independent variable linking to the MIR. The correlations are displayed via double-sided arrows, while the regressions are displayed with a single directional arrow.

The model in Figure 6.3 is the first step of the path analysis and provides the data for further exploration. The second step involves structural equation modeling to test the influence of different model components on cancer fatality with respect to other factors. In other words, this process tests to see if there is a specific sequence of events, represented by the factors, that leads to a better explanation for the outcome.

Figure 6.3. Path model for theoretically grouped factors with the MIR.
Each factor has the possibility of being the initial contributor to the MIR, or falling somewhere along the middle of the path. In total, there are 19 possible pathways to the dependent variable, MIR, and each equation takes all of these into account.

The structural equations created are shown below in equation 6.1. Referring to the pathway diagram (Figure 6.3) for nomenclature, the equations are set up to show the path coefficients based on correlations and coefficients among social characteristics (S), financial and medical access (A), and community and environmental factors (C) that all filter through health and behavioral factors (H) to the MIR (F). Each of the first three path models represent the sum of the direct path coefficient, the adjusted $R^2$ of the model, and the sum of the products from five alternative pathways. The coefficient is represented in the model as a path (p) between two factors, with the destination factor listed first. For example, the path, pFC, would be the coefficient from the community factor (C) to the MIR (F). Referring to figure 6.3, this value would be 0.07. The remainder of the equation represents a sum of all the possible paths originating from the community factor and terminating at the MIR. The last model represents a direct path from health, and is used as a control to determine the extent to which different combinations either improve or detract from the explanatory power of the model.

The construction of the pathways represents the sum of the probabilities of correlations between groups. The pieces of each path were also analyzed to determine the best overall combination of factors. There were two pieces of information driving the pathway tests. The first was the conceptual model, which dictated that the social, financial and environmental factors would influence health, which would in turn influence cancer.
Equation 6.1 - Structural Equation Models for path analysis.

<table>
<thead>
<tr>
<th>Path Model</th>
<th>Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Community and Environmental Path</strong></td>
<td>[ p_{FC} + (p_{HS} \times p_{FH}) + (p_{AC} \times p_{FA}) + (p_{AC} \times p_{HA} \times p_{FH}) + (p_{SA} \times p_{FS}) + (p_{SA} \times p_{HS} \times p_{FS}) + \epsilon_F ]</td>
</tr>
<tr>
<td><strong>Financial and Medical Access Path</strong></td>
<td>[ p_{FA} + (p_{HA} \times p_{FH}) + (p_{SA} \times p_{FS}) + (p_{SA} \times p_{HS} \times p_{FH}) + (p_{CA} \times p_{FC}) + (p_{CA} \times p_{HC} \times p_{FH}) + \epsilon_F ]</td>
</tr>
<tr>
<td><strong>Social Path</strong></td>
<td>[ p_{FS} + (p_{HS} \times p_{FH}) + (p_{CS} \times p_{FC}) + (p_{CS} \times p_{HC} \times p_{FH}) + (p_{AS} \times p_{FA}) + (p_{AS} \times p_{HA} \times p_{FH}) + \epsilon_F ]</td>
</tr>
<tr>
<td><strong>Health and Behavioral</strong></td>
<td>[ p_{FH} + (p_{AH} \times p_{FA}) + (p_{SH} \times p_{FS}) + (p_{SH} \times p_{SA} \times p_{FA}) + (p_{CH} \times p_{FC}) + (p_{CH} \times p_{CA} \times p_{FC}) + \epsilon_F ]</td>
</tr>
</tbody>
</table>

This data led to the construction of pathways that started with one of the three factors and led to health measurements. The second was the PCA grouping, which revealed a combination of social, health and access variables that influenced a majority of the variation in the data and influence over the MIR. All pathways are tested from each point of origin to confirm the model and to determine whether a different formulation may be more appropriate. Of the possibilities, the financial access pathway was found to have the highest path score of 0.7502. Social and health characteristics were the second and third highest, with path scores of 0.7379 and 0.730 respectively. The community factor had the lowest path score of -0.007. The results are shown below in table 6.10.
Table 6.10. Path analysis scores calculated to test the predictive power of each factor as the start of the model. Factor correlations from figure 6.1 are used to derive the path scores.

<table>
<thead>
<tr>
<th>Path Start</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medical and Financial Access</td>
<td>0.7502</td>
</tr>
<tr>
<td>Social Characteristics</td>
<td>0.7379</td>
</tr>
<tr>
<td>Health and Behavioral Characteristics</td>
<td>0.7300</td>
</tr>
<tr>
<td>Community and Environment Characteristics</td>
<td>-0.0070</td>
</tr>
</tbody>
</table>

The purpose of testing the multiple paths was to determine the factor that exerts the most influence when at the start of the path. The path coefficients are all quite close, but the financial access characteristics seem to provide the most explanatory power when filtered through the other factors. Also evident was that the community factor did not add any information to the model. It was still unclear whether the other two factors, health and social, occupied a specific position along a path, or were correlated in the center of the path. To test the seeming lack of contribution from the community factor, another set of path models was constructed, both having access factors first and removing community factors. One path situated the health and social factors in a linear relationship. Another situated them as correlated variables in the center of the model. Figures 6.2 and 6.3 display the model scenarios along with the squared multiple correlation result.
In the case of both alternative models, the construction of structural equations is not necessary. In the previous models, with all factors considered, the squared multiple correlation result is 0.32. In the case of the model in figure 6.2, the result is 0.29, indicating less correlation with the MIR. This model is not a good fit for the data. In Figure 6.3, the correlation is higher, at 0.31, but still not as high as the model with community variables.
Despite the lack of strong correlations to the MIR or other variables, the community factor does still impart some predictive power into the model. Removing the factor reduces the overall fit of the model and is therefore is retained in the final construction of the model. Because of the lower path coefficient, however, it is placed at the end of the path. The final model is presented in Chapter 8, Figure 8.1.
CHAPTER 7

SPATIAL VARIABILITY OF VULNERABILITY FACTORS

This chapter seeks to answer the third and final research question, will people living in different geographic display patterns of disparities based on cancer fatality rates and will the driving factors have different impacts on the MIR dependent on the place? The first part of the spatial analysis involves the creation of hot spot maps for the MIR as well as for each of the theoretical a priori and the PCA grouped variables. Clusters were detected through a Global Moran’s I before a Getis-Ords Gi* statistic was generated for each county. The aim is to identify locations where disparities exist in the US based on clustering of the MIR. In addition to the hot spot analysis, a geographically weighted regression was also run for both the theoretically and PCA grouped factors to determine whether discernable spatial patterns exist amongst the factors and whether the predictive ability of the model (adjusted $R^2$) is improved when spatial weights are taken into consideration.

7.1. Hot Spot Mapping

Utilizing hot spots for the identification of disparity populations should help to resolve many of the issues regarding measurement. The question of whether to use absolute or relative measures of disparity is irrelevant when spatial methods are used. In addition to identifying places experiencing disparities, the complexity of societal
organization can also be displayed through the variations of hot and cold spots among the different factor groups. A hot spot map is constructed using two different spatial statistics. First, a global Moran’s I is calculated for the data. The purpose of this analysis is to detect clusters, which in this case are counties that have either MIR or factor values similar to adjacent counties. Adjacent counties in this research are determined through a fixed-distance band, using a calculated Euclidean distance based on the size of the counties. After clusters of counties with similar values are detected, the second method involved calculating a standardized Getis-Ords Gi* statistic (GiZ score). The GiZ scores are then binned and ranked according to their standard deviation. Hot spots show up red on the map and represent clusters of values more than three standard deviations above the mean. Cold spots show up as blue and represent clusters of values more than three standard deviations below the mean. To better assess the underlying factors, a comparison of the MIR hot spot map is carried out with each of the theoretically grouped factors and then the PCA-grouped factors. The goal is to assess similarities in hot and cold spots for the indicators as they correspond to the MIR. The hot spot map for MIR values is represented in Figure 7.1. Corresponding hot spot maps for each of the factor groupings, both a priori and PCA grouped, are then compared to the MIR map for the visual analysis.

7.1.1. MIR Hot Spots

The first thing that jumps out when looking at the MIR map is a giant hot spot in the Southeast. With the exception of the Florida panhandle and the Atlanta metropolitan area, the region is a cluster of higher cancer fatalities at a 99% confidence level. The fact that cancer incidence and mortality is greatest in the Southeast is already established in
the literature, but the extent and significance of the cluster for fatality still stands out as noteworthy in both its extent and the fact that the high fatality is isolated almost entirely to this region.

Figure 7.1. Mortality to Incidence Ratio (MIR) Hot Spot Analysis. Getis Ords Gi* confidence intervals.

The benefit of this map lies in the identification of areas where the population is experiencing significantly different cancer outcomes. With the exception of the Southeast, only a small area in the upper peninsula of Michigan and the northwest corner of Wisconsin exhibit hot spot clustering, possibly resulting from the lack of data in Minnesota and the Great Lakes. When looking at the clustering of factor values, the correspondence of the major hot and cold spots are compared primarily to the patterns
existing along the east coast as well as some of the major clusters existing along the coast of California and through Colorado, Utah, Idaho and Montana.

7.1.2. Hot Spot Analysis of a Priori Groupings

The four theoretically grouped factors all had correlations to the MIR at statistically significant levels. In addition, each group, with the exception of community variables, revealed a high level of internal consistency in the Cronbach’s alpha test. Both the correlations and the group composition are evident in the hot spot maps the theoretically grouped factors and tend to highlight similar regions corresponding to the MIR. Maps for each of the a priori groups are discussed in the sections below.

7.1.2.1. Social Characteristics

The first theoretical factor, social characteristics, includes variables such as unmarried, single parents, lower religious affiliation and lack of social support along with non-modifiable measures of isolation such as non-English and non-white. The southeast has the highest values for the social factor as indicated by the large hot spot. The spatial extent of the cluster mirrors quite well the results of the MIR map and can be seen below in Figure 7.2.

The relevance of social variables in conjunction with cancer deaths is evident in the large portion of the Southeast coinciding with the MIR data. The social components of the states in this region may be negating the negative impacts of being in a rural area. The significant variables within this factor measure the household structure. Therefore, it can be assumed that regions showing up as cold spots have higher rates of the population
living in a “traditional” family structure with two parents, and that this is an influence on cancer fatality.

Figure 7.2. Social Characteristics Hot Spots. Getis Ords Gi* confidence intervals.

There are a few big exceptions to the alignment of the MIR map with the social characteristics. As opposed to the MIR hot spots, which clustered mainly in the southeast, the social characteristics hot spots exist along most of the US coastline, with the exception of the northeast coast, part of the southwest coast in Texas and the coastline of Oregon and Washington. This pattern stems primarily from the existence of a higher non-white population and a higher percentage of single parent households along the
southeast and southwest coastlines. Figure 7.3 shows a hot spot map for each of these variables.

![Single Parent Households - Hot Spot Map](image1)

![Non-White Population - Hot Spot Map](image2)

Figure 7.3. Single Parent Households and Non-white population hot spot maps. Getis-Ords Gi* confidence intervals.

The disparity based on both of these variables is also evident in the PCA. Non-white and single parent variables represent a majority of the variation within factor two from the model. Also of significance is that the beta value for this factor is the highest from the regression model, indicating a stronger correlation with the MIR.
7.1.2.2. Health and Behavioral Characteristics

The second theoretically grouped factor represents health and behavioral characteristics. This factor contains variables for obesity, smoking, high-risk occupations, low birth weights, lack of exercise, binge drinking, poor general health ranking, and mammograms. Also worth noting is that this factor had the highest correlation with the MIR from the grouped regression model, with a beta value of 0.410. The extent and location of hot and cold spots for this factor, below in Figure 7.4, line up almost perfect spatially with the MIR hot and cold spots, which is evident when comparing the maps.

Figure 7.4. Health and Behavioral Characteristics Hot Spots. Getis Ords Gi* confidence intervals.
The hot spot map corroborates the impact of these characteristics on the MIR as well as the fact that they are highly regionalized. The primary contribution to this factor according the regression analyses and PCA run in previous chapters, would come from the correlation and variance demonstrated in both the obesity rates and poor general health variable. The Atlanta metro area and the Florida peninsula again show up as either less significant (Atlanta is orange as opposed to red), non-significant (central Florida is tan) or as cold spots, which are similar to the trends visible in the MIR analysis. This supports the idea of there also being a spatial correlation between the factors and the outcome measure. Education or income variables would be the most likely underlying variable to explain this relationship, as it also tends to correlate with some of the health behaviors like smoking and exercise. The correlation of education with MIR is 0.381, which is not incredibly strong by itself, but the correlation education has with rural populations is 0.535 and for smoking and exercise it is 0.418 and 0.645, respectively. This could indicate that education is driving health related behaviors, which in turn contribute to the MIR. If this were the case, it would support the evidence already shown in the regression models and the path analysis, which demonstrated a strong correlation between access factors when they were situated at the beginning of the conceptual model.

7.1.2.3. Financial and Medical Access Characteristics

This factor is composed of median household income, education level, income inequality, unemployment rates, mammogram facility access, oncology unit access, uninsured population, the number of doctors and the number of oncologists. The map in Figure 7.5 supports evidence of a relatively strong spatial connection between the
financial indicators and MIR. While the southeast does show up as a hot spot in the same manner as the MIR, there are a number of discrepancies that derail the overall strength of the relationship between the two, however. Two incongruities are visible in Florida and Michigan. Florida shows up as an insignificant cluster in the financial analysis, but as a cold spot in the MIR. Michigan is a hot spot in the financial analysis, but not a part of any significant clustering in the MIR analysis. This may be due to the edge effect of these two states when conducting the cluster analysis. The fact that each of these states has three borders without data may influence the number of counties available for use in the construction of the I-statistic. This same pattern is evident in the southern tip of Texas as well.

Figure 7.5. Financial and Medical Access Characteristics Hot Spots. Getis Ords Gi* confidence intervals.
The hot spot for this factor in the southeast is most likely driven by the education level variable. A hot spot analysis on only this variable, shown below in Figure 7.6, provides spatial evidence of the correlation. There is also evidence of the connection between education and the MIR provided by the regression analysis performed on the PCA grouped factors. As stated previously, education level does seem to influence a number of health variables that are in turn correlated highly with the MIR.

Figure 7.6. Education Hot Spots (percentage of population with a high school diploma). Getis Ords Gi* confidence intervals.
7.1.2.4. Community and Environmental Characteristics

The last theoretically grouped factor is the community and environmental characteristics. It contains variables measuring the number of parks and recreational facilities, the percentage of the county designated as rural, the population growth, density, the per capita pollutant emissions, the number of liquor stores, health food stores and fast food restaurants. Based on research into health care access and many of the issues related to pollution and green spaces, the assumption was that this factor would be every bit as influential as the social and economic factors. As it turns out, the environmental factors, as they are measured in this research, have almost no impact on cancer fatality. The beta value for this group is significantly less than the other three, at only 0.073. The map in Figure 7.7 displays a far more random pattern than exists in the MIR data. There are no large regional clusters evident in the Southeast, making the lack of correlation between the two measures clear. Only the coast of California and the Northeast U.S. have cold spots for community variables that coincide with the MIR. Additionally, the hot spot apparent for this group in the Upper Peninsula of Michigan aligns with a similar set of counties in an MIR hot spot. This factor is the only measure that matches up in this area to explain the fatality rate.

The rural variable is the most significant contributor to the MIR within this factor. Looking at a map showing the distribution of rural and urban areas supports this finding, and this variable tends to cluster more than the community factor in the Southeast. This clustering may be responsible for the higher beta value and of the rural indicator despite its lower correlation with the MIR. Rural areas in the Southeast coincide with the MIR hot spots, but rural areas in the central part of the country do not. Because of this
dichotomy in the correlation, which is dependent on place, the overall predictive ability of the rural indicator is diminished. Without conducting a smaller scale analysis it is difficult to say whether the influence of rural environments is truly of lower consequence to the MIR or whether the influence of the rural indicator changes based on the location.

Figure 7.7 - Community Hot Spots. Getis Ords G* for community environment disparities.

7.1.3. Hot Spot Analysis of PCA Groupings

The hot spot maps for the PCA factors also displayed some interesting trends in the maps. The aspatial data collected for these factors indicates both the variance of the set independent of the MIR and the influence of each factor over MIR. As with the
theoretically grouped factors, the aspatial results still do not reveal the clustering of specific factors or regional differences in the influence of factors over the MIR.

7.1.3.1. PCA Factor 1

The factor one variables from the PCA are a combination of community and financial access indicators. The significant loading indicators include doctors and oncologists, rural population, education level and homeowners. As stated previously, based on the data collected this factor is most likely a measure of access to health care. Looking at the map in Figure 7.8, there are definite spatial similarities evident when compared to the education level map from figure 7.6.

![PCA Factor 1 Hot Spot Map](image)

Figure 7.8. PCA Factor 1 Hot Spots. Getis Ords Gi* for Factor 1 values. Areas in red indicate higher PCA Factor 1 values while blue indicates lower PCA Factor 1 values.
This relationship amongst the other variables within this factor explains why the beta value for the factor 1 is not as high as factor two in the regression model. Just because the variables group together to explain variance does not mean this variance is significant in relation to the MIR. When looking at the map for this factor, it does appear that some relationship may exist between the factor and cancer fatality. Areas in the Southeast corresponding the higher MIRs are highlighted, while areas in the Northeast associated with lower MIRs are also evident.

7.1.3.2. PCA Factor 2

Factor two is defined primarily by the loading of social characteristics along with a few general health indicators. The variables with heavy loadings are non-white, unmarried, single parents, obese, and low birth weight, and the factor had a beta value of 0.359 in the regression against the MIR. Additionally, each of these component variables has a high beta value in the regression analysis, suggestive of a stronger influence on the fatality rates. In the spatial analysis, definite similarities are evident between this factor and the theoretically grouped social factor. Looking at the map in Figure 7.9, there is also a clear pattern that exists between this factor and the MIR. Data from the aspatial analysis suggests that this factor is representative of general health characteristics primarily, which in this case would be obesity and low birth weights. Looking at the spatial patterns, however, it seems as thought this factor correlates more strongly with the social characteristics. In both of these factors, the hot spot extends in the southeast just inside the borders of Louisiana, Mississippi, Alabama, Georgia, South and North
Carolina, and Virginia. This could indicate the influence of state level policies on health and welfare of populations.

Figure 7.9. PCA Factor 2 Hot Spots. Getis Ords G* for Factor 2 values. Areas in red indicate higher PCA Factor 2 values while blue indicates lower PCA Factor 2 values.

The hot spot cluster in the Southeast stands out, as does the cold spot in the Northeast and Central regions of the country. The noticeable patterns that align with the MIR in the spatial context are very important within this group. Even though this second factor does not account for the highest amount of variance in the set, it still does explain over 12% of the total. In addition, it contains four of the top five most significant indicators for the MIR data. This reveals a grouping that is a strong predictor of cancer vulnerability as well as a predictor for cancer disparities.
7.1.3.3. PCA Factor 3

Factor three represents a combination of financial, social and health indicators. There are four variables with significant loadings on the factor, including median household income, income inequality, dependents, and getting a regular mammogram. The very interesting visual pattern in this factor, when looking at the map in Figure 7.10, is the large cold spot in the Atlanta area. This coincides almost perfectly with the gap in the hot spot for the MIR.

![PCA Factor 3 - Hot Spot Map](image)

**Figure 7.10.** PCA Factor 3 Hot Spots. Getis Ords G* for Factor 3 values. Areas in red indicate higher PCA Factor 3 values while blue indicates lower PCA Factor 3 values.

Based on the sets of analyses conducted, the most likely reason for this gap in the cancer fatality hot spot is related to the higher income and education levels present in this
area. Visual patterns are confirmed by the regression on the PCA factors. Factor four still exhibits a good deal of the variance in the set and also has a relatively high beta score, at 0.178. Another interesting finding from this factor showed up in the upper peninsula of Michigan and the Northern part of Wisconsin. In the MIR map, this area represents the only other significant hot spot aside from the one in the Southeast. Factor three has the same hot spot, indicating a possible correlation between the income level and cancer fatalities in that region.

7.1.3.4. PCA Factor 4

Factor four actually has the second highest correlation with the MIR behind factor two in the regression model, but does not account for the same amount of the variance. The correlation to the MIR is not fully explained when looking at the variables that load significantly on the factor. Every variable in the group is a health and behavioral characteristic, including poor health, smoking, and lack of exercise, and none of them have an especially high beta value in the 34-variable regression model. The best description of this factor is related to negative health behaviors. When combined in the factor, these variables account for a higher amount of the correlation than would be expected, suggesting a possible synergistic relationship between the variables. The map in Figure 7.11 clarifies some of this explanatory increase as a result of the spatial similarities between this factor and the MIR hot spots.
Figure 7.11. PCA Factor 4 Hot Spots. Getis Ords G* for Factor 4 values. Areas in red indicate higher PCA Factor 4 values while blue indicates lower PCA Factor 4 values.

The hot spot in the Southeast is very similar in shape and extent to that of the MIR hot spot map, also sharing the same void around Atlanta where the cancer fatality rate is lower. Also evident in the map is the lack of significant hot spots in Florida and the cold spots in the Northeast and West coast. It just so happens that this combination of variables happens to coincide very well spatially, and the explanatory power of this factor is increased as a result.
7.1.3.5. **PCA Factor 5**

Factor five exhibits a smaller hot spot in the southeast as well as some cold spots in the Midwest, but the hot spot in the Southeast is nowhere near the extent of the MIR analysis results. The variance accounted for within factor five comes primarily from lack of access to mammogram and oncological facilities; however these show very weak correlations to the MIR in the aspatial analysis. The map in figure 7.12 shows the existing hot spots in the Southeast along with the cold spots in the Midwest.

Figure 7.12. PCA Factor 5 Hot Spots. Getis Ords G* for Factor 5 values. Areas in red indicate higher PCA Factor 5 values while blue indicates lower PCA Factor 5 values.

This factor may be significant in that it highlights the availability of facilities in the Central U.S. The rural nature of this area, and the rural classification may lead an
over inflation of the access to these facilities because so few people live there. The variables are reliant on the base population as a proxy, and do not provide an indication of how much time and effort it takes a person to get to a facility. This is an important consideration when analyzing data like this and smaller scale analysis would benefit from a drive time analysis or some other proxy with a better representation of access.

7.1.3.6. PCA Factor 6

Primarily community indicators represent factor six in the PCA. Only the parks and liquor store variables load significantly on the component. Following the aspatial analysis, urban areas were mentioned as the probably explanatory variable for this factor, although the hot spot map in figure 7.13 does not support this finding.

The hot spots on this map, in theory, are areas with higher numbers of both parks and liquor stores. Two places on the map suggest that the opposite relationship may be play, however. The map displays cold spot patterns in the Northeast and Central U.S. If this were an urban/rural indicator, these two regions would show up as opposite on the map. A connection between the two factors is only evident in this area and the Northeast; however, so a clear correlation between income, the presence of parks, and lack of liquor stores cannot be made. There also is little to go on in spatial correspondence between the factors and the MIR.
7.1.3.7. PCA Factor 7

Factor seven accounts for the least connection to the MIR among the ten factors, with a beta value of only 0.004 in the regression model and no significance. The map in figure 7.14 does not reveal a much stronger spatial correlation than would be expected. Factor seven is composed of two significant loading variables, religious adherence and high-risk occupations, which are a combination of social and health characteristics. The map seems to highlight areas as hot spots where there would be a higher concentration of careers in higher risk professions like agriculture.
There is no evidence in the literature that these two indicators are correlated with each other, and the analysis conducted in this research does not support a strong correlation either. The relationship between this factor and the MIR appears to be negative based on the visual comparison. The hot and cold spots are reversed from the MIR in many locations along both coasts, but not in the Central U.S. According to the primary loading factors, a hot spot on this map indicates an area with a lower percentage of religious adherents and less high-risk occupations. Considering both of these variables in conjunction with cancer fatality, only the high-risk occupation part makes sense.

Figure 7.14. PCA Factor 7 Hot Spots. Getis Ords G* for Factor 7 values. Areas in red indicate higher PCA Factor 7 values while blue indicates lower PCA Factor 7 values.
Having more high-risk jobs is expected to result in high cancer rates, but not necessarily higher mortality rates. Higher religious affiliation having a positive correlation with cancer fatality definitely does not make sense. The only possible explanation for this factor’s influence may be a relationship to rural areas. Many of the careers classified as hazardous are agricultural, which would most likely set them in rural areas.

7.1.3.8. PCA Factor 8

Factor eight, shown in Figure 7.15, displays a very different spatial relationship than the other factors. In the aspatial regression model, this factor actually has a higher beta value of 0.134. This suggests a stronger correlation between the factor and the MIR.

Figure 7.15. PCA Factor 8 Hot Spots. Getis Ords G* for Factor 8 values. Areas in red indicate higher PCA Factor 8 values while blue indicates lower PCA Factor 8 values.
Analysis of the hot spot map, however, reveals almost no clustering of the factor across the country. The factor eight hot spots are represented primarily by only a single significant loading factor, TRI releases. Looking at the map, it seems as though all counties are roughly equivalent with the exception of some coastal regions. The cold spots in these areas are representative of higher TRI releases and most likely oil refineries, based on their locations. Compared spatially to the MIR map, this map of factor eight seems to have no predicative ability.

7.1.3.9. PCA Factor 9

The correlation between the MIR and factor nine is relatively weak, and the spatial association between the two is weak as well, as visible in figure 7.16.

Figure 7.16. PCA Factor 9 Hot Spots. Getis Ords G* for Factor 9 values. Areas in red indicate higher PCA Factor 9 values while blue indicates lower PCA Factor 9 values.
The significant loading variables in this factor are population growth and density. In the PCA component matrix, the density variable is negatively correlated. This means that a hot spot on the map is indicative of high population growth and lower density. There are only a few areas identified as hot spots, and most of them do seem to coincide with areas of lower cancer fatality as well. This relationship may explain some of the predictive power seen for the factor.

7.1.3.10. PCA Factor 10

Factor 10 is represented primarily by the lack of health food accessibility variable in the analysis. Surprisingly, this seems to coincide with many of the urbanized areas, as evident in Figure 7.17 below.

Figure 7.17. PCA Factor 10 Hot Spots. Getis Ords G* for Factor 10 values. Areas in red indicate higher PCA Factor 10 values while blue indicates lower PCA Factor 10 values.
The only areas where this factor lines up spatially with the MIR are in the Southeast. Aside from this area, there seems to be a negative spatial correlation with the MIR. Highly urbanized areas appear to correspond unexpectedly with the lower density of healthy food. This is an unexpected relationship, and would indicate that the health food density is inversely correlated with cancer fatality. In the regression and correlation models; however, this is not corroborated.

7.2. Bivariate Moran’s I

While each of the hot spot maps is useful for the identification of significant clusters for both the MIR and each of the groups, it does not measure or help to visualize the extent of overlap between the MIR and the groups. In order to accomplish both of these tasks, a bivariate Moran’s I is employed using GeoDa software. In this analysis the MIR is compared spatially with each group to identify the relationship between the two. Four categories are created in the cluster analysis: high MIR/high group value, high MIR/low group value, low MIR/high group value, and low MIR/low group value. In addition to this, a Moran’s I value is created.

Queen contiguity to the second degree is used in GeoDa to derive a k-nearest neighbors spatial weights matrix. This neighbor selection method was chosen to accommodate the variation in county size between the East and West coast of the US. Queen contiguity takes both the counties with shared borders as well as those with shared vertices, effectively creating a circle around the county for analysis. Going to the second degree encompasses the next level of counties around the original ring. This ensures that even counties on borders or coastlines will still have at least four nearest neighbors in the analysis. The data obtained through the bivariate Moran’s I analysis reveals different
patterns of spatial correspondence between the each group and the MIR, shown below in Figures 7.18 – 7.20.

7.18. Bivariate Moran’s I maps comparing MIR with each of the theoretical groups (\(\alpha = .05\))

7.2.1. Results from Theoretical Groupings Bivariate Moran’s I

Looking first at the theoretical groupings and their correspondence to the MIR, a few themes emerge. The first is the number of significant relationships as well as the type of relationship between the group and the MIR. The most significant spatial correlations are found in the map with Health and Behavioral Characteristics, where only 1,107 of the counties did not show up as significant in the analysis. Social characteristics
correspond to the MIR in many of the counties, with only 1,305 not having a significant relationship of some kind. Financial and Medical Access Characteristics also have significant correlations among less than half of the counties. Community and Environmental characteristics were the only grouping to not have a significant correlation in more than half of the US counties. The Moran’s I values mostly confirm the cluster maps as well. Table 7.1 shows the results for each of the bivariate Moran’s I analyses.

With the exception of the Financial and Medical Access Characteristics, the visual analysis corresponds to the I-values. The likely reason for this is due to the clustering of the groups. The social characteristics are more diffuse throughout the country, making the correlations look more pervasive. Part of the Moran’s I value is determined by clustering of similar relationships, however, and the Financial and Medical Access Characteristics have a slightly higher level of clustering between the counties with similar MIR and group values.

Table 7.1. Bivariate Moran’s I Test results for Theoretical Groupings (α=.05)

<table>
<thead>
<tr>
<th>Theoretical Grouping</th>
<th>Bivariate Moran’s I</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social Characteristics</td>
<td>0.178828</td>
</tr>
<tr>
<td>Health and Behavioral Characteristics</td>
<td>0.345607</td>
</tr>
<tr>
<td>Financial and Medical Access Characteristics</td>
<td>0.216706</td>
</tr>
<tr>
<td>Community and Environmental Characteristics</td>
<td>0.017489</td>
</tr>
</tbody>
</table>
This brings up the topic of the relationship between groupings and the MIR, whether the MIR is positively or negative correlated to the group values. Counties with either high-high or low-low correlations represent positive relationships, while negative relationships are represented by high-low and low-high relationships. Both the health and the access characteristics maps have higher percentages (greater than 70%) of counties that correlate positively with the MIR, while the social and community characteristics have a more even split between the negative and positive correlations.

7.2.2. Results from Inductive Groupings Bivariate Moran’s I Analysis

The same analysis was run for the ten inductive grouped factors in order to compare both the extent of spatial overlay as well as the nature of the relationship between each factor and the MIR. Figures 7.19 and 7.20 display the ten maps created using GeoDa and serve as part of the analysis.

Because of the difficulty in directly comparing each factor map, a table was created as well to help with the analysis. (Table 7.2) This helps to see both the percentage of counties with a significant overlap between the MIR and factors as well as being able to determine the percentage of counties within each factor that have positive or negative correlation.
In the bivariate maps created for the inductive groupings, a few of the factors reveal patterns of correspondence with the MIR. The number of counties with a significant correlation between the factors and the MIR appears relatively consistent amongst most of the maps. Factor 4 had the highest amount of overlap with the MIR, with factors 2 and 7 coming in as the only others with greater than 50% of the counties having a significant correlation to the MIR. This is also evident when looking at the maps for these factors. With regards to the type of relationship evident, factor 4 also displayed the highest number of positively correlated counties. Factors 1, 2, and 3 also had more than 60% of the counties with a positive correlation.
Figure 7.20. Bivariate Moran’s I MIR maps comparing inductive groups 6-10 (α=.05)

The spatial distribution of the overlap is also worth noting. Much of the high-high correspondence in the first four factors takes place in the Southeast, where the higher MIR rates are clustered. This spatial overlap of factors with the MIR in the Southeast is a good indication of potential interactions. Factors one and four represent primarily health, behavioral, and access indicators, while factors one, two and three are primarily social indicators. The findings here support those of the regression analysis on the inductive groupings, where social and health indicators accounted for much of the influence on the MIR.
Table 7.2. Bivariate Moran’s I Results for Inductive Groupings (α = .05)

<table>
<thead>
<tr>
<th>Factor</th>
<th>Significant Counties Count</th>
<th>Total Counties in Analysis</th>
<th>Significant Counties (%)</th>
<th>High-High Counties</th>
<th>Low-Low Counties</th>
<th>Positive Factor-MIR Correlation (%)</th>
<th>Low-High Counties</th>
<th>High-Low Counties</th>
<th>Negative Factor-MIR Correlation (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1,326</td>
<td>3,259</td>
<td>40.7</td>
<td>439</td>
<td>449</td>
<td>67.0</td>
<td>326</td>
<td>112</td>
<td>33.0</td>
</tr>
<tr>
<td>2</td>
<td>1,866</td>
<td>3,259</td>
<td>57.3</td>
<td>457</td>
<td>835</td>
<td>69.2</td>
<td>180</td>
<td>394</td>
<td>30.8</td>
</tr>
<tr>
<td>3</td>
<td>1,580</td>
<td>3,259</td>
<td>48.5</td>
<td>507</td>
<td>561</td>
<td>67.6</td>
<td>259</td>
<td>253</td>
<td>32.4</td>
</tr>
<tr>
<td>4</td>
<td>2,093</td>
<td>3,259</td>
<td>64.2</td>
<td>672</td>
<td>813</td>
<td>71.0</td>
<td>291</td>
<td>317</td>
<td>29.0</td>
</tr>
<tr>
<td>5</td>
<td>1,279</td>
<td>3,259</td>
<td>39.2</td>
<td>419</td>
<td>294</td>
<td>55.7</td>
<td>394</td>
<td>172</td>
<td>44.3</td>
</tr>
<tr>
<td>6</td>
<td>805</td>
<td>3,259</td>
<td>24.7</td>
<td>69</td>
<td>385</td>
<td>56.4</td>
<td>160</td>
<td>191</td>
<td>43.6</td>
</tr>
<tr>
<td>7</td>
<td>1,790</td>
<td>3,259</td>
<td>54.9</td>
<td>340</td>
<td>489</td>
<td>46.3</td>
<td>529</td>
<td>432</td>
<td>53.7</td>
</tr>
<tr>
<td>8</td>
<td>756</td>
<td>3,259</td>
<td>23.2</td>
<td>265</td>
<td>129</td>
<td>52.1</td>
<td>305</td>
<td>57</td>
<td>47.9</td>
</tr>
<tr>
<td>9</td>
<td>1,039</td>
<td>3,259</td>
<td>31.9</td>
<td>167</td>
<td>266</td>
<td>41.7</td>
<td>232</td>
<td>374</td>
<td>58.3</td>
</tr>
<tr>
<td>10</td>
<td>1,387</td>
<td>3,259</td>
<td>42.6</td>
<td>300</td>
<td>399</td>
<td>50.4</td>
<td>380</td>
<td>308</td>
<td>49.6</td>
</tr>
</tbody>
</table>

7.3. Geographically Weighted Regression

Following the hot spot map analysis for each of the factors and the MIR, a geographically weighted regression (GWR) was run for each of the groups, both theoretically grouped and methods. A GWR is used, as opposed to an Ordinary Least Squares (OLS), due to the proven variations in the data from the hot spot analysis. These regressions were run using an adaptive kernel type determination method utilizing the Akaike Information Criterion (AIC). This resulted in 133 nearest neighbors being used for the theoretical groupings and 234 nearest neighbors for the inductive groupings. The projection used to run this analysis was a North American Albers Equal Area Conic in order to minimize distortion of the data. The purpose of the GWR models was to examine the two grouping methods using the spatially weighted data and to confirm the
presence of all explanatory variables in the model. The benefits gained from using a GWR analysis are primarily in the accounting for some of the spatial autocorrelation that exists amongst the data. It is clear from looking at the maps that there are spatial patterns in the data and that not all of the relationships between the factors and the MIR are similar, nor are they consistent through space. The downside to the GWR, much like the other analyses completed in this study, is in the masking of local patterns and lack of predictive ability outside of the study region. None of the relationships visible at the US scale can be assumed to hold true at a local level, nor can they be applied to data in other countries. A separate analysis and evaluation of relationships would have to be conducted in order to apply data at different scales or in different locales.

The first GWR model run was for the theoretically grouped factors. The results of the model are shown in Table 7.3. There are two numbers important for comparative purposes in this table, the AICc value and the adjusted $R^2$. The AICc value will be used to compare the model fit of this regression to the PCA grouped GWR. A difference of greater than three indicates a model that is better. The adjusted $R^2$ can be compared to the non-spatial regression run previously on the same factors and the MIR.

Table 7.3. Results of GWR analysis on theoretically grouped factors. MIR: dependent variable.

<table>
<thead>
<tr>
<th>VARIABLE NAME</th>
<th>VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neighbors</td>
<td>133</td>
</tr>
<tr>
<td>Residual Squares</td>
<td>1,312.968</td>
</tr>
<tr>
<td>AICc</td>
<td>6,498.067</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.525</td>
</tr>
<tr>
<td>$R^2$ Adjusted</td>
<td>0.457</td>
</tr>
</tbody>
</table>
Comparing the adjusted $R^2$ values of this model to the aspatial regression reveals some improvement. The adjusted $R^2$ of the aspatial regression on the theoretically grouped factors was 0.324, whereas this model produced an adjusted $R^2$ of 0.457. This improvement provides further evidence that spatial patterns are a better predictor of cancer outcomes than using only aspatial methods. In addition the output feature class of the GWR below in Figure 7.21, reveals a random pattern of residual values. This indicates a well-specified regression model and that all independent variables are accounted for in the analysis.

![Geographically Weighted Regression - Theoretical Groupings](image)

Figure 7.21. Geographically Weighted Regression output feature class residuals for theoretically grouped factors.
The second GWR analysis was run using the ten PCA grouped factors. All of the same data was collected for this model and used both to compare with the aspatial regression as well as with the theoretically grouped factors in a spatial regression. Table 7.4 displays the pertinent information for the model results, and figure 7.22 is a map of the residuals for the model.

Table 7.4. Results of GWR analysis on PCA grouped factors. MIR: dependent variable.

<table>
<thead>
<tr>
<th>VARIABLE NAME</th>
<th>VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neighbors</td>
<td>234</td>
</tr>
<tr>
<td>Residual Squares</td>
<td>1,253.380</td>
</tr>
<tr>
<td>AICc</td>
<td>6,481.217</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.547</td>
</tr>
<tr>
<td>$R^2$ Adjusted</td>
<td>0.470</td>
</tr>
</tbody>
</table>

The map shows a random pattern for the residuals similar to the results from the theoretically grouped factor GWR. This again indicates as well constructed spatial regression and the presence of all variables. A comparison between the two GWR models, however, reveals that the theoretically grouped factors perform slightly better than the PCA grouped factors predicting the MIR. This is based on an AICc value of 6498 for the theoretically grouped factors and 6481 for the PCA grouped factors. A difference of more than three is indicative of a significant difference in the models and a higher number is better.
7.4. Summary of Visual Analysis

The important information learned from both the hot spot analysis and the GWR is that spatial relationships are evident in the data set and that the spatial correlations between the factors and the MIR add to the predictive power of the model. The mapping results could have just as easily turned up a completely random pattern showing no clustering at all throughout the country. With the lower $R^2$ values of the aspatial models, this result was very likely. What this analysis proved was that spatial clustering methods using health outcome data for cancer is an effective technique for identifying disparities. Clear regional patterns emerged on most the maps, and an improvement in predictive
ability of the data was evident when using the GWR as opposed to an aspatial regression model.

There were some interesting patterns that showed up in the MIR, factors and the correspondence between them as well. One of the patterns that showed up in each of the three analyses was the “hole” around the Atlanta area. The area showed up in contrast to the surrounding Southeastern patterns in many of the results. The cold spot that showed up in the MIR map corresponds to high-high overlaps visible in the health and behavioral characteristics bivariate map as well as in the financial and medical access characteristics. There are similarities between the social characteristics and the MIR as well, but there seems to be more confinement to the Southeast. Similar correspondence can also be detected between the MIR and PCA factor one, factor three, and factor four. Each of these factors share strong similarities to the access and health characteristics. Based on these findings, it seems as though there is a connection between access and health, as well as a connection between these two factors and the probability of incurring negative cancer outcomes.

Another interesting pattern exists in the Northeastern U.S., where a sizeable cold spot exists for the MIR. This cold spot corresponds positively in the bivariate map with health and behavioral characteristics and financial and medical access characteristics, while have some overlap with community and environmental characteristics. Factors one, two, and four appear to have more significant low-low overlaps with the MIR. In this case, the health and access are still apparent, however, factor two is more associated with social indicators. This suggests that mechanisms making a population more vulnerable may not be the same mechanisms that make other populations less vulnerable.
The combination of characteristics is not a static entity in either a spatial or temporal sense.

Using a spatial analysis, in this case, reveals the need for a sub-regional analysis for health research and that different characteristics will have impacts on population health dependent on the location. The lack of predictive ability seen in each of the factors is the result of this spatial variation. In all likelihood, analysis of the same factors at a smaller scale would reveal a much higher correlation since this spatial variation would be removed.
CHAPTER 8

CONCLUSIONS

There were three main goals of this research, defined by the research questions posed. The first was to identify drivers of vulnerability to cancer deaths in the U.S. The second was to find the broad-based factors accounting for most of the spatial variability, and by extension, disparities in cancer deaths. The third and final goal is to create an accurate assessment of spatial patterns in the data that reveal populations exposed to more of the cancer burden. The research was set up to systematically answer each of these questions through a variety of aspatial and spatial methods.

8.1. Research Question 1

Research question 1 asked, “what are the predominant socio-spatial factors driving vulnerability to cancer deaths in the United States?” As expected, the answer is a complicated mix of characteristics that do not maintain consistent relationships with the outcome measure. Despite this complexity, however, some predominant trends did show up. There are a combination of variables contributing to both the social and the health and behavioral characteristics of a place. These variables include the obese population, non-white population, and the number of single parent and unmarried households.

Three major pieces of evidence support this conclusion. First, the beta values for both the proportion of single parent and unmarried households in the county were among
the highest in the 34-variable regression model and had significance as well. Second, in
the PCA analysis, the percentage of non-white population in the county also correlated
with the proportion of unmarried and single parent households in the county, loading
heavily on the second factor. This indicates that these three indicators account for much
of the variability within the set, and are most likely responsible for a good deal of the
cancer disparities in a county. In addition, the regression model run on the PCA grouped
factors revealed the highest correlation to be between factor two and the MIR. These
social variables, along with two health variables in this factor, accounted for a beta value
of 0.359 in this model. The third part of the analysis in support of the conclusion was the
spatial patterning of the indicators that lined up exceptionally well with the MIR. The
clustering of the social characteristics factor appeared to be very similar spatially to that
of the MIR. Factor two from the PCA model also lined up very well with the MIR.

The groupings were rearranged in the beginning of the research, after the research
questions were generated. At this time the thought was that education would translate
better to the social characteristics in the model. After further research, education level
was moved to the financial and medical access characteristics due to the more likely
correlation between doctors, hospital services, and education levels. Despite the removal
of this variable from the social characteristics, it still did seem to exert a great deal of
influence on the model.

This was one of the more interesting findings, actually, and supported the
eventual rearrangement of the model. The evidence for this influence came from two
places. The first was the direct correlation between the education level and the MIR. In
addition to this direct connection, the education variable also showed up in both the
theoretically grouped and the PCA grouped models as a significant contributor to other access variables.

8.2. Research Question 2

Research question two asks, “Which of the broad based factors accounts for most of the spatial variability in cancer outcomes?” The answer to research question two is more complex than the first. The goal of this portion of the research was to determine which of the groupings accounted for the most spatial variation. The first, and simpler answer came from looking at the PCA groupings. Factor one in this analysis accounted for the most variance and is composed of education level, renters, doctors, oncologists, rural areas, and fast food, each loading significantly. While the health and behavioral characteristics did have the highest correlation to the MIR as measured in the regression model, they were not responsible for the variability, which would in turn lead to disparities.

The answer to this research question is complicated; however, when considering the spatial correspondence to cancer outcomes. This part of the question is better answered by looking at the PCA results; the results of PCA grouped regression and the hot spot maps for the factors. While the first factor in the PCA model does account for most of the variability in the data set, the beta value for this factor is not nearly as high as that of factor two, which is represented by the non-white population, single parents, unmarried, obesity, and low birth weight variables. The hot spot map for factor two also shares a lot more commonalities to the MIR hot spot map than factor one when conducting a visual inspection. Between these two pieces of evidence, it appears as
though a combination of social and health characteristics are responsible for a majority of
the spatial variation in cancer outcomes.

The correspondence of the factors to the MIR was also evident in the analysis of
the bivariate Moran’s I maps and tables. The strength of the positive spatial correlations
between factors two, three and four are evident in both, supporting the connection
between the access characteristics, health characteristics and social characteristics. The
PCA factors each represent a blend of indicators, each also significant in the 34-variable
regression model from RQ1. Of the three, factor four has the highest amount of overlap
and the most counties with a positive correlation to the MIR. This suggests health and
behavioral characteristics, primarily a combination of smoking and lack of exercise, are
the most influential factor in the spatial variability of cancer outcomes in the U.S.

8.3. Research Question 3

Research question 3 asked, “will people living in different geographic areas
possess different driving factors for cancer mortality and will the factors consistently
display the same directionality of impact?” There are a number of very interesting spatial
patterns that turned up in this section of the research. The goal of the spatial analysis was
to both establish spatial correlations between the MIR and each of the groups in addition
to demonstrating an added strength when considering spatial variability.

The difference in influence of factors between locations was very clearly
illustrated in the visual analysis and through the results of the GWR. There were a
multitude of variations in pattern between each of the factor groups and the MIR.
Alignment would indicate that the MIR correlated with the factor similarly. Instead, the
maps proved that a couple of scenarios are possible. One is that each factor presents
differently in a place, and will therefore have a different impact on cancer fatality rates. The other possibility is that the interaction of factors in a place leads to different manifestations of cancer outcomes, again dependent on the place. In either instance, the prediction of cancer fatality is not easily accomplished through the use of assumed variable relationships.

The urban/rural divide did potentially show up on the MIR map, demonstrating a possible connection between this variable and the vulnerability to death from cancer. The significance of this relationship stems not from the actual rural/urban divide, but more likely from the presence of other variables that coincide with urban or rural areas as stated previously. The influence of variables may differ depending on whether they are present in an urban or rural location. This proves further evidence that smaller scale research is important and may reveal an answer to this question.

The Bivariate Moran’s I analysis contributed significantly to assessing the spatial overlap between the factors and the MIR, allowing for a quantitative analysis of these relationships. If the relationships between factors and the MIR were consistent across the county, the bivariate maps would be primarily two colors representing either high-high or low-low relationships. Instead, there is a mix of both these positive correlations along with a nearly equal number of negative correlations (i.e. high-low). While some of the factors do have a higher level of overlap with the MIR, there does not appear to be a consistent or predictable trend in the correlations among any of the factors and the MIR. Of the factors, financial and medical access characteristics and health and behavioral characteristics are the most closely correlated to the MIR and share the highest spatial
correspondence, supporting the conclusion that these two factors would most consistently display a similar directionality of impact.

8.4. Research Summary and Significance

The final research question is the most significant contribution of this research to the field of cancer disparities. The goal of this research question was to determine whether vulnerability factors for cancer were consistent throughout the United States. The expectation, as stated in the first hypothesis, was that drivers would exert different forces on cancer fatalities dependent on the place. This hypothesis was derived because of the known difficulty in establishing consistent drivers of both cancer incidence and cancer fatality in other studies. The lack of consistency was confirmed in this research as well through the result of a low adjusted $R^2$, even when considering thirty-four variables known to associate with cancer outcomes. If each of these variables possessed a consistent relationship to cancer, the predictive power of the set would have been significantly higher. There are many possible explanations for the low predictive ability of the model, including missing variables, improper spatial scale for the analysis, and the influence of spatial variations in the data. As a result of this finding, two recommendations can be made with respect to future cancer research. The first is that multiple variables must be considered in the search for drivers of disparities. A single variable cannot be expected to explain the cause for either fatality rates or disparities between groups. Path analysis holds the greatest potential in this field, revealing the most likely root cause amongst a group of variables. The second recommendation is that researchers refrain from any assumptions of applicability from one case study to another. While this is not an implicit assumption of research in the field, there do seem to be a
number of studies that analyze similar constructs or use only specific sets of variables. The results of research conducted here suggest that, unless within very close spatial proximity to another study, no assumptions are made as to which variables should be used to predict cancer outcomes or how the variables will associate with the outcomes.

8.5. Research Applications

As stated previously, the significance of these research findings results from the establishment of the complex interaction of variables dependent on place. Prior to the accommodation of spatial variations, it is important to first consider how the variables interact with each other in a place. This involves reassessing the conceptual model from the beginning. Figure 8.1 displays the revised concept model for cancer disparities based on this research. Following the diagram is a description of each component and the interactions.

After looking back through the results of the research it became apparent that establishing predetermined relationships among the factors would likely have a negative impact on future endeavors to assess cancer disparities. The idea behind running a spatial analysis was to, in part, establish the variability of relationships between societal characteristics and the MIR. Having shown this variability to exist at the U.S. scale, it would not make any sense to then create a conceptual model implying specific relationships.
An inductive approach does appear to be the best approach to determine the strength and order of relationships that best describe the societal constructs leading to cancer disparities in a place. Based on this thought, the best model is one in which all variables are assumed as equal contributors. A correlation between the variables is also assumed to exist.

The new model does not propose any starting point in the determination of cancer disparities. While there is some evidence from the path analysis to support placing financial and medical access first in the model, there is no guarantee that the same relationship will occur at different scales. In addition, there is a good deal of evidence presented in the spatial analysis that suggests different relationships between the factors and the MIR that are dependent on place. Taking this into consideration, a more suitable

Figure 8.1. Revised concept map based on results presented.
model should refrain from the use of directionality. Instead of having a definitive starting point, each of the factors in this model is displayed as having a direct influence on cancer disparities. The strength of the connection is not implied because it may change.

A benefit of this model for future research is its ability to be operationalized. There is a great deal of potential to use this model to test high and low MIR counties against each other, for example. The relationships among factors and the MIR can be assessed to determine specific alterations that may exist, leading to the disparities. Using this proposed model to establish strengths of relationships and to run path analyses, models specific to a place could be created. This research can serve as an example of this model specification. At the U.S. level, there were specific interactions evident among the factors and the MIR that could be used to create a model explicit to this data set and geography. An example of this directional model is presented in Figure 8.2.

![Diagram of Conceptual Model created using U.S. county-level data.](image)

The relationships found in this U.S. level research support a certain level of model directionality as well as specific correlations between the factors. Financial and medical
access characteristics would most likely form the base of this U.S. specific model along with social characteristics as determined through the structural equation modeling. There are strong factor correlations between the health and behavioral characteristics, the financial and medical access characteristics, and the social characteristics that suggest a close relationship among the three. This relationship was confirmed through analysis of factor correlations as well as through visual analysis using hot spot maps of the factors.

The community and environmental characteristics shift in the new model as well and act as a filter through which the cancer disparities are ultimately defined. The evidence for the movement of this factor to the last position in the model comes primarily from the spatial analysis. The hot spot maps revealed a pattern in the MIR that coincided with the urban/rural divide, which is the most significant indicator within the community factor. Despite this spatial correlation the community factor did not predict the MIR well at all. Not all rural areas corresponded to higher cancer fatalities even though they are easily distinguishable on the map. Based on this evidence, it appears as though the community factor is highly dependent on the characteristics present in the other three factors.

Testing the function of this urban/rural divide as a filter for the other factors is the next step in better defining the drivers of cancer disparities. The characteristics of a place have proven to influence the patterns of interaction between variables that lead to cancer fatalities. This precludes any future large-scale research and demands a shift to regional analysis. Urban and rural areas should be studied independent of each other with no a priori assumptions as to the influence of variables on cancer outcomes. A PCA run on all known indicators of the MIR will identify the variables most likely contributing to the
disparities and inform a path analysis to derive the underlying cause of the higher or lower fatality rate. Regional analysis based on the urban/rural divide should lead to two different sets of models, each modified to reflect the influence of variables in that place. The two resulting models can be compared in order to clearly establish unique patterns in each location. In the end, the goal is to eradicate the disparities in cancer deaths and ensure that these rates drop equitably for all groups. Identifying a location where rates are significantly different is only half of the battle, and should only be useful to establish the causes of the difference. If the reasons change based on location, as was proven, the use of a place-based model is essential in order to first establish the significant variables and then predict how they will influence the outcomes.

8.6. Caveats

There are some limits to note in this research based on the available data and the geography. First, the county level analysis conducted in this research will not reflect trends in smaller geographic areas. There is a lot of potential variability within counties, and this certainly causes smaller scale trends to be hidden in the data. Running smaller scale analysis in the future would allow for this problem to be circumvented. Data availability for many of the variables was responsible for the county level measure. The goal from the beginning was to establish a method and model that would allow scaling. This would be dependent on data availability and the needs of the local or regional medical staff or other individuals concerned with cancer vulnerability.

The use of all-cancer rates also presents some potential issues in the interpretation of results. Not all cancers have similar incidence or mortality rates, and therefore certain cancer types may dominate the data. For example, breast cancer and prostate cancer
higher incidence rates, yet relatively low mortality rates. Because of this relationship, these two cancers will drive many of the trends visible in this research. As stated previously, however, the goal of this research was to establish overall cancer disparities, regardless of cancer type. Breast cancer and prostate cancer can be fatal if not treated properly or diagnosed too late. Smaller scale studies may benefit from breaking out specific cancers, but this should be done after an initial analysis of the cancer prevalence in the area and the societal characteristics present.

There exists an assumption in this research that geographic variability stems directly, and solely, from the factors being studied. In reality, there may be an element of chance, or general variability present in the outcomes that has no connection to the factors. This variability potential is greater in this study due to the size of the data set and the spatial extent of the data, and is not accounted for.

Another limitation of this research lies in the use of some transformed variables. Although the data is intended as a comparative analysis, looking at the data values with respect to the scores of others in the set, the process of transformation can cause some issues in the interpretability of the data. It was deemed more important to have all linear variables and be able to conduct a linear regression model than to adapt the model to the presence of non-linear variables and lose a good deal of the predictive power.

Lastly, the inference of a causal relationship between the identified drivers and the outcomes does not guarantee its actual existence. There are a multitude of possibilities that could exist. Aggregating up to the county level can allow for ecological fallacies, meaning that many of the county level assumptions may not apply to smaller scale patterns. Individual-level assumptions cannot be made from this research. Also,
inherent in geographic research of this variety is the assumption of treatment in a local area. Mobility of the population cannot be fully accounted for, and people could easily cross county lines for medical services. Data inaccuracies can also play a role in the accuracy of the research. Registry data, for instance, relies on proper diagnosis codes and accuracy in reporting could create problems.

Another major consideration that must be accounted for in future research is the cancer type. This study used the all-cancer incidence and mortality rates for each county. The predominant cancer types in each county undoubtedly sway the data analysis. Breast cancer may have higher prevalence in a specific area and result in an inaccurate assumption that mammograms are highly correlated, for instance. Adding cancer type to the analysis will be pertinent to the regional analysis. It is likely that certain regions will possess one or two predominant cancer types. In this case, a model should be constructed for each, as it is probable that different variables will have different contributions dependent on the cancer type in addition to the place.
REFERENCES


Fay, M.P. 1991. “Approximate confidence intervals for rate ratios from directly
standardized rates with sparse data.” *Communications in Statistics-Theory and
Methods.* 28(9): 2141-2160.


Environment Better for Health? If So, Why?” *Annals of the Association of

Census: Religious Congregations & Membership Study,” *Association of
Statisticians of American Religious Bodies.*


Relevant to Healthy People 2010 Cancer-Related Objectives,” in *NCI Cancer
Surveillance Monograph Series, Number 6.* Bethesda, MD: National Cancer
Institute.

Incidence Ratios to Illustrate Racial and Sex Disparities in a High-Risk

Cancer.* 131(12): 2724-2732.

Helquist, M. 2010. "Effectiveness of population-based service screening with
mammography for women ages 40 to 49 years." *Cancer.* 118(4): 1169-1169.

Hewett, K. and I. Burton. 1971. *The Hazardousness of Place: A Regional Ecology of
Damaging Events.* Toronto: University of Toronto Press.


# APPENDIX A

## DETAILED DATA SOURCES

Table A.1: Detailed variable list with computations and source URL.

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Source URL</th>
<th>Table Name/Variable ID</th>
<th>Computation (if any)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIR – Incidence Rate</td>
<td><a href="http://www.statecancerprofiles.cancer.gov/incidencerates/index.aspx">http://www.statecancerprofiles.cancer.gov/incidencerates/index.aspx</a></td>
<td>Each state individually pulled.</td>
<td>Denominator of MIR. Each state is pulled using all cancer sites, all races, and for both sexes.</td>
</tr>
<tr>
<td>MIR – Mortality Rate</td>
<td><a href="http://www.statecancerprofiles.cancer.gov/deathrates/deathrates.html">http://www.statecancerprofiles.cancer.gov/deathrates/deathrates.html</a></td>
<td>Each state individually pulled.</td>
<td>Numerator of MIR. Each state is pulled using all cancer sites, all races, and for both sexes.</td>
</tr>
<tr>
<td>Median Household Income</td>
<td><a href="http://factfinder2.census.gov/tables/servlet/producview.xhtml?pid=ACS_09_5YR_S1903&amp;prodType=table">http://factfinder2.census.gov/tables/servlet/producview.xhtml?pid=ACS_09_5YR_S1903&amp;prodType=table</a></td>
<td>S1903</td>
<td>No computation. 2009 ACS 5-year estimate for all U.S. Counties.</td>
</tr>
<tr>
<td>Income Inequality</td>
<td><a href="http://www.countyhealthrankings.org/resources">http://www.countyhealthrankings.org/resources</a></td>
<td>GINI</td>
<td>GINI Index – Income inequality Range (0-1). This is already calculated in the data table</td>
</tr>
<tr>
<td>Unemployed</td>
<td><a href="http://factfinder2.census.gov/tables/servlet/producview.xhtml?pid=ACS_09_5YR_S2301&amp;prodType=table">http://factfinder2.census.gov/tables/servlet/producview.xhtml?pid=ACS_09_5YR_S2301&amp;prodType=table</a></td>
<td>S2301</td>
<td>No computation. This variable represents the percentage of the civilian population unemployed.</td>
</tr>
<tr>
<td></td>
<td>(2000) <a href="http://factfinder2.census.gov/tables/servlet/producview.xhtml?pid=DEC_00_SF1_P001&amp;prodType=table">http://factfinder2.census.gov/tables/servlet/producview.xhtml?pid=DEC_00_SF1_P001&amp;prodType=table</a></td>
<td>2000 – P001</td>
<td></td>
</tr>
<tr>
<td>Renters</td>
<td><a href="http://factfinder2.census.gov/tables/servlet/producview.xhtml?pid=ACS_09_5YR_S2501&amp;prodType=table">http://factfinder2.census.gov/tables/servlet/producview.xhtml?pid=ACS_09_5YR_S2501&amp;prodType=table</a></td>
<td>S2501</td>
<td>No computation. Percentage of renter-occupied housing units.</td>
</tr>
<tr>
<td>Race-Non-white</td>
<td><a href="http://factfinder2.census.gov/tables/servlet/producview.xhtml?pid=ACS_09_5YR_B02001&amp;prodType=table">http://factfinder2.census.gov/tables/servlet/producview.xhtml?pid=ACS_09_5YR_B02001&amp;prodType=table</a></td>
<td>B02001</td>
<td>[(Total Pop. – White only Pop.) / Total Pop.] x 100 = % non-white population</td>
</tr>
<tr>
<td>Religious affiliation</td>
<td><a href="http://www.thearda.com/Archive/Files/Downloads/RCMSCY10_DL2.asp">http://www.thearda.com/Archive/Files/Downloads/RCMSCY10_DL2.asp</a></td>
<td>All Denominations/Groups</td>
<td>No computation. County level congregation membership per 1,000 in total population.</td>
</tr>
<tr>
<td>Category</td>
<td>URL</td>
<td>Code</td>
<td>Description</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>----------------------------------------------------------------------</td>
<td>------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td></td>
<td>tview.xhtml?pid=ACS_09_5YR_S1201&amp;prodType=table</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single-parent household</td>
<td><a href="http://factfinder2.census.gov/tables/servlet/pages/produc">http://factfinder2.census.gov/tables/servlet/pages/produc</a></td>
<td>B11001</td>
<td>(Male householder + female householder) / total</td>
</tr>
<tr>
<td></td>
<td>tview.xhtml?pid=ACS_09_5YR_B11001&amp;prodType=table</td>
<td></td>
<td>households = % single parent households</td>
</tr>
<tr>
<td>Number of dependents</td>
<td><a href="http://factfinder2.census.gov/tables/servlet/pages/produc">http://factfinder2.census.gov/tables/servlet/pages/produc</a></td>
<td>S1101</td>
<td>(% Families with &gt;1 over 60) + (% Families with &gt;1 under18) = % of households</td>
</tr>
<tr>
<td></td>
<td>tview.xhtml?pid=ACS_09_5YR_S1101&amp;prodType=table</td>
<td></td>
<td>with at least one dependent</td>
</tr>
<tr>
<td>Educational Attainment</td>
<td><a href="http://factfinder2.census.gov/tables/servlet/pages/produc">http://factfinder2.census.gov/tables/servlet/pages/produc</a></td>
<td>S1501</td>
<td>No computation. % population 25 or over with a high school diploma or</td>
</tr>
<tr>
<td></td>
<td>tview.xhtml?pid=ACS_09_5YR_S1501&amp;prodType=table</td>
<td></td>
<td>equivalency.</td>
</tr>
<tr>
<td>Language isolation</td>
<td><a href="http://factfinder2.census.gov/tables/servlet/pages/produc">http://factfinder2.census.gov/tables/servlet/pages/produc</a></td>
<td>S1602</td>
<td>No computation. % households with no person over 14 proficient in English.</td>
</tr>
<tr>
<td></td>
<td>tview.xhtml?pid=ACS_09_5YR_S1602&amp;prodType=table</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parks per thousand</td>
<td><a href="http://www.arcgis.com/home/item.html?id=6618daa42a2ab4b13846d4b37b29dc60">http://www.arcgis.com/home/item.html?id=6618daa42a2ab4b13846d4b37b29dc60</a></td>
<td>North America Parks</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Download data file along with county shapefile with population counts</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>in 2010. Clip NA park feature and get count of parks per county.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Export to excel and calculate parks per thousand.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(Number of parks / county population) x 1000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recreation Facilities</td>
<td><a href="http://www.countyhealthrankings.org/resources">http://www.countyhealthrankings.org/resources</a></td>
<td></td>
<td>No computation. Per capita count of recreational facilities in a county.</td>
</tr>
<tr>
<td></td>
<td>U7VAKhbl1usY</td>
<td></td>
<td>No computation. Index for livability of area based on climate factors.</td>
</tr>
<tr>
<td>Environmental hazards</td>
<td><a href="http://iaspub.epa.gov/triexplorer/tri_release.chemical">http://iaspub.epa.gov/triexplorer/tri_release.chemical</a></td>
<td>Total On-site and total off-site emissions.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total amount of emissions from TRIs in county per capita. Choose</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>year 2009 and Total of all on site and off site emissions for all</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>chemical classifications. Not adjusted for population.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rural population</td>
<td><a href="http://factfinder2.census.gov/tables/servlet/pages/produc">http://factfinder2.census.gov/tables/servlet/pages/produc</a></td>
<td>H2</td>
<td>(Number of housing units classified rural / Total housing units) x 100 = %</td>
</tr>
<tr>
<td></td>
<td>tview.xhtml?pid=DEC_10_SF1_H2&amp;prodType=table</td>
<td></td>
<td>Rural housing</td>
</tr>
<tr>
<td>Particulate Matter Days</td>
<td><a href="http://www.countyhealthrankings.org/resources">http://www.countyhealthrankings.org/resources</a></td>
<td>Particulate Matter Days</td>
<td></td>
</tr>
<tr>
<td></td>
<td>No computation. Value is already calculated in the data file.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Number of days the particulate latter exceeded safe limits.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ozone Days</td>
<td><a href="http://www.countyhealthrankings.org/resources">http://www.countyhealthrankings.org/resources</a></td>
<td>Ozone Days</td>
<td></td>
</tr>
<tr>
<td></td>
<td>No computation. Value is already calculated in the data file.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Number of days the level of ozone exceeded safe levels.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Liquor Store Density</td>
<td><a href="http://www.countyhealthrankings.org/resources">http://www.countyhealthrankings.org/resources</a></td>
<td>Liquor Store Density</td>
<td></td>
</tr>
<tr>
<td></td>
<td>No computation. Value is already calculated in the data file.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Density of liquor stores per square mile in the county</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Number of computation. Value is already calculated in the data file.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Number of fast food restaurants per 1000 population. This is from</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>the 2011 data file and calculated based on 2008 data.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High risk occupation</td>
<td><a href="http://factfinder2.census.gov/tables/servlet/pages/produc">http://factfinder2.census.gov/tables/servlet/pages/produc</a></td>
<td>CB0900A1</td>
<td>NAICS Codes Filter – 00(all), 1151, 21, 22, 23, 3122, 313, 32, 562, 72, 811</td>
</tr>
<tr>
<td></td>
<td>tview.xhtml?pid=BP_2009_00A1&amp;prodType=table</td>
<td></td>
<td>(High Risk sector workforce (sum) / Total workforce in all sectors (00)) x</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>100 = % High Risk occupations.</td>
</tr>
<tr>
<td>Category</td>
<td>URL</td>
<td>Description</td>
<td>Calculation</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>----------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Health Food Access</td>
<td><a href="http://www.countyhealthrankings.org/resources">http://www.countyhealthrankings.org/resources</a></td>
<td>Health Food Access</td>
<td>No computation. Value is already calculated in the data file. Percentage of zip codes in county with healthy food options</td>
</tr>
<tr>
<td>Population density</td>
<td><a href="http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_09_5YR_G001&amp;prodType=table">http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_09_5YR_G001&amp;prodType=table</a></td>
<td>G001 – Geographic Identifiers - Land area in square miles</td>
<td>(County population / Land area in square miles) = Number of people per square mile.</td>
</tr>
<tr>
<td>Smoking</td>
<td><a href="http://www.countyhealthrankings.org/resources">http://www.countyhealthrankings.org/resources</a></td>
<td>Percentage of smokers</td>
<td>No computation. Value is already calculated in the data file. Percentage of population (&gt;18) who smoke</td>
</tr>
<tr>
<td>Alcohol</td>
<td><a href="http://www.countyhealthrankings.org/resources">http://www.countyhealthrankings.org/resources</a></td>
<td>Percentage of binge drinkers</td>
<td>No computation. Value is already calculated in the data file. Percentage of population (&gt;18) who consume &gt; 5 (male) or 4 (female) alcoholic beverages at a time</td>
</tr>
<tr>
<td>Exercise</td>
<td><a href="http://www.countyhealthrankings.org/resources">http://www.countyhealthrankings.org/resources</a></td>
<td>Exercise</td>
<td>No computation. Value is already calculated in the data file. Percentage with less than daily recommended exercise</td>
</tr>
<tr>
<td>Obesity</td>
<td><a href="http://www.countyhealthrankings.org/resources">http://www.countyhealthrankings.org/resources</a></td>
<td>Obese Population</td>
<td>No computation. Value is already calculated in the data file. Percentage of population (&gt;20) with BMI &gt; 25</td>
</tr>
<tr>
<td>Mammography Units</td>
<td><a href="http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfm/mqsa.cfm">http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfm/mqsa.cfm</a></td>
<td>Mammogram facilities</td>
<td>Each state data pulled individually and geocoded. Mobil units assigned to county registered in. County counts created in ArcMap. (Number of mammogram units in county / county population) x 1,000</td>
</tr>
<tr>
<td>“poor” general health</td>
<td><a href="http://www.countyhealthrankings.org/resources">http://www.countyhealthrankings.org/resources</a></td>
<td>Rank health as poor</td>
<td>No computation. Value is already calculated in the data file. Percentage of population ranking health as “poor”</td>
</tr>
<tr>
<td>Low Birth Weight</td>
<td><a href="http://www.countyhealthrankings.org/resources">http://www.countyhealthrankings.org/resources</a></td>
<td>Low birth weight</td>
<td>No computation. Value is already calculated in the data file. Percentage of live births with babies weighing less than 5 pounds.</td>
</tr>
<tr>
<td>No social support</td>
<td><a href="http://www.countyhealthrankings.org/resources">http://www.countyhealthrankings.org/resources</a></td>
<td>No social support</td>
<td>No computation. Value is already calculated in the data file. Percentage of population reporting no social support</td>
</tr>
<tr>
<td>Number of doctors</td>
<td><a href="http://ahrfrhsa.gov/download.html">http://ahrfrhsa.gov/download.html</a></td>
<td>MD and DO counts</td>
<td>(Number of practicing doctors / county population) x 100,000 = Doctors per 100,000 population</td>
</tr>
<tr>
<td>Number of internal MDs</td>
<td><a href="http://ahrfrhsa.gov/download.html">http://ahrfrhsa.gov/download.html</a></td>
<td>Internal Medicine Doctors</td>
<td>(Number of practicing internists / county population) x 100,000 Number of internal Medicine DRs per 1,000 population</td>
</tr>
<tr>
<td>Hospitals with oncology service</td>
<td><a href="http://ahrfrhsa.gov/download.html">http://ahrfrhsa.gov/download.html</a></td>
<td>Hospitals with oncology services per 1,000 population</td>
<td>Hospitals with oncology services per 1,000 population</td>
</tr>
</tbody>
</table>