GEOGRAPHIC CONCENTRATIONS OF PROSTATE CANCER INCIDENCE IN TEXAS AND THEIR RELATIONSHIP TO SOCIOECONOMIC CONDITIONS

THESIS

Presented to the Graduate Council of
Southwest Texas State University
in Partial Fulfillment of the Requirements

For the Degree

MASTER OF SCIENCE

Department of Geography

By

J. Gaines Wilson, B.S.

San Marcos, Texas
May, 2002

ACKNOWLEDGEMENTS

First and foremost, I would like to express my gratitude to my thesis research supervision committee, which included Drs. F. Benjamin Zhan (Chair), Jean Brender, and Deborah Bryan. Dr. Brender introduced me to the world of epidemiology and carefully guided me in interpreting the results of the research. Dr. Bryan provided much-needed assistance in the development of my proposal and with issues relating to geographic information systems. Dr. Zhan was instrumental in helping me in every phase of the project from conception to delivery. One of the most delightful and intelligent professionals I have ever worked with in business or academia, his encouragement and kind advice helped me bring this project from fruition to successful completion.

In my research and academic activities, several other people were of tremendous help. Dr. Lawrence Estaville guided me through the development of the theoretical framework for this research. Xuwei Chen and Guangyu Wu provided data manipulation support that was critical to the core of this research. Emily Manderson, a fellow graduate student, provided the finishing touches on almost every version of the manuscript and furnished most of the moral support necessary to complete this thesis.

Finally, I would like to thank my parents, William A. "Art" and Kathy "KiKi" Wilson. Without their financial support, true-north guidance, and years of patience with my varied directions in life, I would more than likely be the focus of research rather than an investigator.

CONTENTS

Chapter Page

LIST OF TABLES

LIST OF ILLUSTRATIONS

ABSTRACT

Prostate cancer is a dangerous and elusive disease. The disease is the most diagnosed cancer and the second highest leading cause of death from cancer in Texas men. Although prostate cancer is highly prevalent in male populations worldwide as well as in Texas, the environmental risk factors for the disease are relatively unknown. For these reasons, research about the environmental risk factors for prostate cancer is of great importance and urgency.

This thesis addresses four research questions: (1) Are there any statistically significant spatial clusters of prostate cancer incidence in Texas at the census tract level?; (2) is there any statistically significant association between prostate cancer incidence clusters and socioeconomic status at the census tract level in Texas?; (3) is there any significant association between prostate cancer incidence clusters and rural place of residence in Texas?; and (4) is socioeconomic status a significant risk factor for prostate cancer incidence in Texas at the census tract level? To answer these questions, several research objectives were met.

The first objective was to investigate the geographic distribution of prostate cancer incidence in Texas at the census tract level. This objective was achieved using a spatial scan statistic cluster test developed by Kulldorff (1997). The scan statistic was applied to prostate cancer incidence data from the Texas Cancer Registry of the Texas Department of Health and population data from the U.S. Census for three race-ethnicity

categories (white, black, and Hispanic), and four age groups (18-24, 25-44, 45-64, 65+). A statistically significant most likely cluster was detected in north San Antonio and 23 significant secondary clusters were observed in other sections of Texas.

The second objective was to investigate socioeconomic status and place of residence as risk factors for prostate cancer incidence clusters and prostate cancer incidence rates. Logistic regression and Poisson regression analyses were employed to reach this objective. Median household income and census tract education were chosen as the indicators of socioeconomic status and people per square mile was the indicator of rural or urban areas. High income, high education, and urban place of residence were all found to be significant predictors of counties containing prostate cancer incidence clusters. Further, high income, high education, and urban place of residence were found to be risk factors for prostate cancer incidence rates independent of clusters. In more precise terms, higher socioeconomic status and living in an urban area make one more prone to developing prostate cancer.

This research contributes to the established literature of environmental spatial analysis and spatial epidemiology. For the first time, this research demonstrated that: (1) there is a significant cluster of prostate cancer incidence in northern San Antonio at the census tract level in Texas, and (2) socioeconomic status and urban place of residence are risk factors in *both* prostate cancer incidence clusters and prostate cancer incidence rates at the census tract level in Texas. Medical researchers and public health planning officials may benefit from this research by using the results found here to further focus prostate cancer research.

CHAPTER 1

INTRODUCTION

Prostate cancer is a dangerous and elusive disease. This research will apply spatial epidemiology methods to search for unusually concentrated areas of prostate cancer in Texas. This research will also evaluate certain environmental exposures as risk factors for developing prostate cancer.

The remainder of this chapter is divided into five sections. The first section will discuss the significance of the research. Research questions are the subject of the second section, and the limitations of the research are discussed in the third section. The fourth section introduces major concepts related to the research and the fifth section serves as an overview and reader's guide to the rest of this thesis.

Significance of the Research

This research is significant for several reasons, the first being the seriousness of the disease. Prostate cancer is a dangerous disease in that it can lead to death. Cancer of the prostate poses a serious health risk to the male populations of most western countries (Hsing and Devesa 2001; Dijkman and Debruyne 1996). It is the most commonly diagnosed non-skin cancer in most western countries and is the second leading cause of death in United States men following only lung cancer (Hsing and Devesa 2001; Ross

and Schottenfeld 1996). This year alone, cancer of the prostate will be the most commonly diagnosed cancer and the second leading cause of death from cancer in both the United States and Texas male populations (Hanchette and Schwartz 1992; Hsing and Devesa 2001; Kelada et al. 2000; Ross and Schottenfeld 1996; Shibata and Whittemore 1997; Texas Department of Health 2001; Yu, Harris, and Wynder 1998). In 2001, approximately 32,000 U.S. men died from the disease and about 180,000 men were newly diagnosed (Hsing and Devesa 2001; Johns Hopkins Oncology Center 2001). Among Texas men, prostate cancer remains the most common type of cancer (Texas Department of Health 2001). Each year 12,000 new cases are diagnosed and 2,000 men die in the state (Texas Department of Health 1999; Texas Department of Health 2001). Prostate cancer presents a significant health problem in Texas and deserves further investigation.

The second reason why this research is significant is because the environmental risk factors for prostate cancer are relatively unknown. The disease is elusive. Although prostate cancer is prevalent in male populations worldwide, the risk factors for cancer of the prostate are mostly unknown, from both a spatial and an etiological perspective (Hsing and Devesa 2001; Potosky, Feuer, and Levin 2001). Unlike many cancers, where a cause-effect relationship can be established (i.e. cigarette smoking and lung cancer), there is no evidence that any specific environmental or social factor increases the risk of developing prostate cancer (American Cancer Society 2001; Hsing and Devesa 2001; Meade and Earickson 2000; National Institutes of Health 2001; Ross and Schottenfeld 1996). Prostate cancer should be investigated because we do not yet know what environmental risk factors cause the disease.

The final reason why this research is significant is because of the lack of current literature and studies in Texas on prostate cancers cluster at the census tract level as well as prostate cancer studies in Texas that investigate socioeconomic status (SES) as a risk factor. A thorough library, Internet and medical database search did not produce any previous literature on statewide prostate cancer and SES census-tract level studies conducted in Texas. Although studies have been conducted in other states, the Texas Department of Health Cancer Data request point person was not aware of any Texas SES-prostate cancer studies in the literature (Risser 2001a). The disparity of historical research on this topic in Texas is further justification of its significance.

Research Questions

This paper addresses four research questions:

1. Are there any statistically significant spatial clusters of prostate cancer incidence in Texas at the census tract level?
2. Is there any statistically significant association between prostate cancer incidence clusters and socioeconomic status at the census tract level in Texas?
3. Is there any significant association between prostate cancer incidence clusters and rural place of residence in Texas?
4. Is socioeconomic status a significant risk factor for prostate cancer incidence in Texas at the census tract level?

The research questions above will be answered using the following four null hypotheses:

1. There is no significant clustering of prostate cancer incidence in Texas at the census tract level.
2. Socioeconomic status is not significantly correlated to prostate cancer incidence clusters in Texas at the census tract level.
3. Rural place of residence is not significantly correlated with prostate cancer incidence clusters in Texas at the census tract level.
4. Socioeconomic status is not a significant risk factor for prostate cancer incidence in Texas at the census tract level.

In order to test these null hypotheses, several research objectives will be met. First, cluster analysis will be performed using the prostate cancer incidence data at the census tract level. Next, logistic regression analysis will be used to detect any associations present between socioeconomic status and prostate cancer incidence clusters at the census tract level. Logistic regression analysis will also be performed to determine any statistically significant associations between rural place of residence and prostate cancer incidence clusters. Finally, a Poisson regression will be performed to test if socioeconomic status is a risk factor for prostate cancer incidence at the census tract level.

Limitations of the Research

This research is limited from several perspectives. The limitations of the research include limitations of the source data, spatial limitations, and temporal limitations.

The cancer incidence case data is limited in that it cannot accurately represent all known cases for the study period. Many men with prostate cancer have not yet been

diagnosed with the disease and thus, the disease remains latent and unreported (Hsing and Devesa 2001). The population data from the census is also not entirely accurate due to under-reporting in some age-race-ethnicity categories.

To draw useful conclusions in this study, as in all spatial epidemiologic studies, certain spatial associations need to be made. Data on socioeconomic conditions is not available on a case-by-case basis. In order to match cases to a certain socioeconomic group, each individual case is associated with the average socioeconomic conditions for the tract in which they live in. Of course, census tracts are not homogenous in their socioeconomic conditions. One person within a census tract could be in the upper quartile of education and income while a close neighbor within the same tract could possibly be in the lowest quartile of education and income.

Temporal factors limit the accuracy of the research as well. The assessment of prostate cancer incidence in a diverse population requires that different age-race-ethnic groups be compared to one another separately over each year of the study period, 1990-1997. The 1990 census data for age-race-ethnicities in Texas counties is available, but 2000 data for age-race-ethnicities at the tract level was not yet available at the time the research was conducted. This presented a difficulty in accurately estimating the population growth for each tract-year in each year of the research. The years 1991 through 1997 were therefore approximated using a linear growth rate based on the Texas male population growth rate between 1990 and 2000. This approximation may have skewed the annual population estimates of each age-race-ethnicity.

Major Concepts Related to the Research

Several concepts are important to understand the nature of the research, going forward. These concepts include epidemiology, spatial epidemiology, and some of the terms associated with these concepts.

Epidemiology

Epidemiology is the study of the distribution and determinants of disease frequency in human populations, or, the study of the distribution and determinants of health related states or events in specific populations and the application of this study to control health problems (Brender 2001a, Last 2001). This research is of an epidemiological nature in that it is concerned with both the distribution and determinants of prostate cancer in Texas. Several of the basic terms used in epidemiology are defined in the following sections.

Rates and Standardized Incidence Ratios

Rates are one of the most important concepts in epidemiology. Rates are defined as the measure of a frequency of occurrence of a phenomenon (Brender 2001c). That focus phenomenon in this research is prostate cancer incidence.

Incidence describes the extent that people within a population who do not have a disease will develop that disease during a specific amount of time (Timmreck 1994). In more precise terms, incidence is the new cases of a disease over a specified period of time divided by the population at risk (Brender 2001b). For prostate cancer, incidence rates are expressed as a rate of cases per 100,000 population (TDH 2001).

Prostate cancer, like many other diseases, affects different age-race-ethnic groups in varying degrees of severity (Dale et al. 1996; Hoffman et al. 2001). Population groups also differ by age, race, and ethnicity. To compensate for these differences, rates can be expressed by age, sex, and race. Age-race-ethnic specific rates are defined as the number of cases in a certain age-race-ethnic group over a certain time period divided by the population at risk in that age-race-ethnic group over the same time period (Brender 2001c). Rate adjustment is used to account for rate differences in ages, races, and ethnicities across populations. Because prostate cancer incidences vary by age-race-ethnicity and because census tracts in Texas differ in their age-race-ethnicity distribution, it was necessary to use rate adjustment in this thesis.

The expected number of cases in each census tract for each age-race-ethnicity category was calculated using the indirect method of adjustment. The standard population in this thesis includes all men of the twelve age-race-ethnicity categories in the state of Texas from 1990-1997. The indirect method of adjustment applies the age-race-ethnicity rate from the standard population to the age-race-ethnicity population distribution of a specific tract to arrive at the expected cases. These expected cases are summed to obtain total expected cases in a tract. The observed and expected cases are used in the calculation of standardized incidence ratios.

Standardized incidence ratios (SIR) were computed for use as the dependent variable in the regression models. The standardized incidence ratio is the observed number of cases divided by the expected number of cases. SIRs for each age/race/ethnicity category will be determined by using the indirect method of rate adjustment. Before computing the SIR of a census tract, the total number of observed

incidences and expected incidences must be calculated. General formulas for rates, incidence rates, adjusted rates and standardized incidence ratios are given in figure 1.1.

FIGURE 1.1

Formulas for Rates, Incidence Rates, Adjusted Rates, and Standard Incidence Ratios (Source: Brender 2001a, 2001b)

$$\text{Rate} = \frac{\text{Number of Cases}}{\text{Population at Risk Over Time Period}} \times 100{,}000$$

$$\text{Incidence Rate} = \frac{\text{Number of New Cases Over Period}}{\text{Average Population Over Period}} \times 10^{N}$$

where N is dependent on the convention being used,

N=5 in most prostate cancer incidence rate expressions

$$\text{Age - Race - Ethnicity Specific Rate} = \frac{\text{Number of Cases in Age Group}}{\text{Population of Age Group In Time Period}} \times 100{,}000$$

$$\text{Standarized Incidence Ratio}_i = \frac{O_i}{E_i} \times 100$$

where i represent a region within the study area, O_i denote the total observed cases within a study region and E_i the total number of expected incidences within a study region

Testing and Latency

Prostate cancer incidence rates have risen sharply over the last decade in western countries. This rise is not necessarily due to more people developing the disease, but because of the increasing awareness from more accurate tests, such as the prostate-specific antigen (PSA) test (Crocetti, Ciatto and Zappa 2001). Prior to the PSA test, testing by other methods was highly inefficient and rare. In the late 1970s, the incidence rate for all U.S. males was estimate to be 19.74 per 100,000 (Macdonald and Heinze 1978). By the early 1990s, incidence rates experienced a spike to over 100 per 100,000 in all U.S. males, mostly due to the introduction of PSA testing in the late 1980s. According to Statistics, Epidemiology and End Results (SEER) data from the National Cancer Institute, white male incidence increased 70% from 1980 to 1990 (Crocetti and Zappa 2001). This trend peaked in the early 1990s (Figure 2.1) when PSA testing reached a saturation point in the population. Since then, the rate of growth has stabilized (National Caner Health Statistics 1999). Even with the introduction of the PSA test, prostate cancer remains highly latent in male populations: 50% of men over the age of 70 are living with undiagnosed (latent) tumors (Hsing and Devesa 2001). Texas also experienced a sharp rise in prostate cancer incidence rates in the late 1980s. This growth curve leveled off and a slight decline in the trend has been recorded since the early 1990s due in part to effective PSA screening (Texas Department of Health 2001).

TABLE 1.1

Age Adjusted Incidence Rates (per 100,000) for
Selected Cancer Sites According to Race

Group	1990	1991	1992	1993	1994	1995
White	133.0	169.1	188.3	163.4	140.0	129.8
Black	173.3	223.3	256.9	270.6	245.7	211.6

Source: National Center for Health Statistics. 1999. *Health, United States, 1999 with health and aging chartbook.* Hyattsville, MD: National Center for Health Statistics.

Reader's Guide to this Thesis

The remainder of this thesis is organized as follows. Chapter two summarizes the literature of relevance to the research. Chapter three outlines the methods used in the research. Chapter four focuses on the research and results of the cluster analysis and chapter five reviews the research and results of the regression analyses. Chapter six summarizes the primary conclusions, the contributions of the research and the future opportunities for further research.

CHAPTER 2

LITERATURE REVIEW

The existing literature on prostate cancer is quite extensive. This chapter consists of several sections that summarize the literature relevant to this research. The first section provides and overview of the framework of theory within which the research fits. The second section outlines the etiology of prostate cancer and section three describes the risk factors for prostate cancer.

Theoretical Framework

Research reported in this thesis fall within the general theoretical framework of spatial analysis, a grand tradition in geography. In addition, the research also contributes to human ecology theory.

Spatial analysis concerns itself with the variations in the localization and distribution of a significant phenomena or group of phenomena (Holt-Jensen 1988). The origins of spatial data analysis (SDA) date back to the Quantitative Revolution in geography and regional sciences in the early 1960s (Fischer 1999; Zhou 2000). The notion of transforming maps to analyze relationships between objects was given an early impetus by Tobler and Bunge, both of whom supported a use for mapping in other spaces other than those just physical (Gatrell 1983). Finally, in 1968, the book *Spatial Analysis*

provided a solid foundation of readings from which the discipline would develop (Martin and James 1993). Over the next few years, many SDA techniques were developed in fields like such as archaeology, ecology, epidemiology, geography, geology, and urban and regional planning (Bailey 1994; Zhou 2000). Spatial analysis is most closely associated with positivist explanation in geography, because it usually deals with formal modes of spatial organization and assumes objective, certain knowledge of spatial arrangements and space-time processes (Gatrell 1983, 4). This thesis adopts a positivist approach to spatial analysis in its pursuit of objective risk factors that might contribute to prostate cancer incidence using modern tools of spatial analysis, namely geographic information systems (GIS) and spatial statistics.

The focus of this thesis is on a specific branch of spatial analysis, spatial epidemiology. Spatial epidemiology is defined as the analysis and description of spatial and spatial-temporal distributions of disease data (Elliot et al. 2000; Haining 1998; Zhou 2000). One way to understand cancer etiology is to consider geographic variations in human cancer rates (Higginson 1983). For almost 150 years, spatial epidemiology has been used to solve perplexing problems in public health. British physician John Snow utilized maps in what is now the classical epidemiological approach (Snow 1854). By comparing rates of cholera in London to water supplies, he concluded that an "impurity" in the water was associated with cholera cases (Selvin 1996). Although current studies remain true to some of Snow's original methods, modern day tools available to medical geographers and spatial epidemiologists like GIS have increased the power of studies by many orders of magnitude. This project will use standard methods in the field of medical geography to investigate the research questions with the goal of ultimately contributing to

the understanding of the locations of clusters of prostate cancer incidence in Texas and the environmental risk factors associated with prostate cancer.

A secondary theory related to this research, human ecology theory, is concerned with "the spatial and sustenance relationships in which human beings are organized ... in response to the operation of a complex of environmental and cultural forces" (Knox 1982, 59). Human ecology theory originated from the work by several urban sociologists in Chicago led by Robert Park (Park, Burgess, and McKenzie 1925). In their original work, *The City,* they define four ecological classifications of communities that are similar in nature to the socioeconomic classifications used in this research (Park, Burgess, and McKenzie 1925). Human ecological theory highlights the reciprocal interaction between individuals and their environment. One concept within human ecology theory is the nested environment model. The nested environments include four categories: micro, meso, exo, and macro (UNT 2001). Figure one shows the nested environments. This research will focus on the interactions between social status and human health at the meso and exo environments by examining social variables at the census tract level. Social area analysis, a subset theory of social differentiation, is used in this thesis to define social rank and urbanization by age, race and ethnicity.

Figure 2.1

Nested Environments of Ecological Theory (Source: UNT 2001)

Etiology of Prostate Cancer

Unlike some other cancers, prostate cancer does not produce any immediate, obvious symptoms until the late stages of the disease (Johns Hopkins Oncology Center 2001). Eventual symptoms include frequent or difficult urination, painful ejaculation, blood in the urine or semen, and frequent pain in the lower back, hips, or upper thighs (Johns Hopkins Oncology Center 2001). Several screening tests over the past 15 years have made significant strides in detecting prostate cancer in its earlier stages. One of the most common of those tests is the Prostate-Specific Antigen (PSA) test (Crocetti, Ciatto and Zappa 2001). PSA is a serine protease that is prostate-specific but not prostate-cancer specific. Men with PSA values above (4.0ng/ml) will have cancer only 1/3 of the time or less (Miller and Torkko 2001). The President of Southwest Texas State University, Dr. Jerome Supple, discovered that he had a prostate cancer in the early stages through the results of a PSA test. His cancer is now in remission (The Daily University Star 1998).

One way to gauge the severity of the onset and progression of a disease is to examine the five-year survival rate of the cancer. When detected early, like most diseases, prostate cancer has a high five-year survival rate. Figure 2.1 shows the improvement in five-year U.S. survival rates over two decades. The improvement in survivability is usually attributed to the introduction of the PSA test and greater health screening awareness among the public (Crocetti, Ciatto and Zappa 2001).

TABLE 2.1

Five-Year Relative Prostate Cancer Survival Rate Percentages for Selected Cancer Sites by Race

Group	1974-79	1980-82	1983-85	1986-88	1989-94
Whites	70.0	74.5	77.7	85.2	95.1
Blacks	60.5	64.7	64.0	69.2	81.2

Source: National Center for Health Statistics. 1999. *Health, United States, 1999 with health and aging chartbook.* Hyattsville, MD: National Center for Health Statistics.

Risk Factors

A plethora of risk factors have been suggested for prostate cancer, but its etiology is still relatively unknown (Ross and Schottenfeld 1996; Meade and Earickson 2000; American Cancer Society 2001; Hsing and Devesa 2001). Common factors for many types of cancers, such as smoking, do not significantly increase the risk of developing prostate cancer (Kelada et al. 2000). A few of the known risk factors for prostate cancer are age, race, ethnicity, and familial history (Carter, Carter, and Isaacs 1990; Dale et al. 1996; Dayal, Polissar and Dahlberg 1985; National Institutes of Health 2001). Risk factors that may influence prostate cancer incidence and mortality include socioeconomic status and rural place of residence (Baquet et al. 1991; Carter, Carter, and Isaacs 1990; Cella et al. 1991; Dale et al. 1996; Dayal, Polissar, and Dahlberg 1985).

Prostate cancer is a major age-related malignancy (Djikman and Debruyne 1996; National Institutes of Health 2001). The disease is rare before age 40, and then incidence rates double for each subsequent decade of life. This age-specific incidence curve for prostate cancer has a steeper slope than for any other cancer, and prostate cancer has the highest age-adjusted incidence rates of any malignancy in U.S. men 65 years and older, 1457.7 (black males) and 932.2 (white males) per 100,000 population (National Institutes of Health 2001; Ross and Schottenfeld 1996). Over 71% of prostate cancer cases occur in men age 65 and over and approximately 92% of prostate cancer deaths occur in the age group 65 years and older (Hanchette and Schwartz 1992; National Institutes of Health 2001). Accordingly, it is important to adjust for age when calculating and comparing rates of prostate cancer.

The most apparent epidemiologic observation about prostate cancer is the striking differences in incidence rates among racial/ethnic groups. A comprehensive review of the literature reported that all studies measuring socioeconomic status (SES) and race found that race was a significant risk factor, even when SES was not considered (Dale et al. 1996). This difference in racial incidence can be as high as *forty*-fold between Chinese men and U.S. black men (Hsing and Devesa 2001). In the United States, the three groups generally studied are blacks, whites and Hispanics. There is significant disparity of cancer outcomes (incidence, survival and mortality) for African Americans, white Americans, and Hispanic Americans (Hardy and Hargreaves 1991). Among diverse male population groups in the United States, black and white men have the highest prostate cancer incidence rates and higher mortality rates than all other races (National Institutes of Health 2001). African Americans have twice the risk of non-Hispanic whites for developing advance-stage prostate cancer. (Hoffman et al. 2001; Mebane, Gibbs, and Horm 1990). Further, in every age group, at every clinical stage, and in every histological grade, black U.S. men have much higher rates than U.S. whites (Dale et al. 1996). The reasons for the increased risk of U.S. black men are still unknown (Hsing and Devesa 2001). Hispanics generally have a lower incidence of prostate cancer than both non-Hispanic whites and blacks (Hoffman et al. 2001, Liu et al. 2001, Shibata and Whittemore 2001). In Texas, these racial and ethnic trends play out as they do on the national level (Texas Department of Health Cancer Registry 2001). In 1997, non-Hispanic whites, blacks and Hispanics had respective age-adjusted incidence rates (per 100,000) of 56.6, 71.1 and 36.9 (Texas Department of Health Cancer Registry 2001).

Socioeconomic Status as a Risk Factor for Prostate Cancer

Studies in the literature have been inconclusive about the role of socioeconomic status (SES) in determining risk for prostate cancer (Dale et al. 1996). The factors used to determine SES vary widely. However, most studies state that at least education and income should be considered in measuring SES (Baquet et al. 1991; Dale et al. 1996; Dayal and Chiu 1982; Dayal, Polissar, and Dahlberg 1985; Earnster et al. 1978; Hoffman et al. 2001; McWhorter et al. 1989; Polednak 1990; Ross et al. 1979; Yu et al. 1988). Table 2.2 provides a summary of selected studies, including the factors used to calculate SES in each study.

Social class has been shown to be inversely associated with mortality (Hardy and Hargreaves 1991). People of lower socioeconomic status have lower life expectancies and higher mortality rates in almost all causes of death (Deonandan et al. 2000). Some diseases, such as lung cancer, are traditionally associated with the poor (Elliott et al. 1996). Others, such as breast cancer, have more often afflicted the affluent. (Meade and Earickson 2000). While population studies do suggest that there may be genetic components to some prostate cancers, there are likely to be numerous environmental components to prostate carcinogenesis and prevention (Brawley and Barnes 2001). For example, African Americans have 4 times the incidence of native Africans, but share the highest rates worldwide with Jamaicans (Hsing and Devesa 2001, Jones 2001). Because Jamaican, African, and U.S. blacks are similar in their genetic makeup, these differences suggest a geo-environmental risk factor. The literature supports that race is a risk factor for developing prostate cancer (Dale et *al.* 1996). However, it is not yet known how this

TABLE 2.2

Cancer and Socioeconomic Status Studies in the Literature

Author(s)	SES Factor(s)	Site(s) Studied	Findings
Hoffman et al. 2001.	• employment status • household income • insurance status • marital status	Prostate	Blacks found to have twice the risk of non-Hispanic whites. SES accounts for 15% of the increases relative risk in blacks.
Dale et al. 1996.	• income • education	Prostate and others	Review article of 176 studies. Concluded that at least income and education should be used in all future SES prostate cancer studies.
Vijayakumar et al. 1992.	• median income	Prostate	Blacks have higher risk.
Baquet et al.1991.	• education • family income • population density	>100	Low SES explains much of the excess cancer burden. For prostate cancer, blacks have higher risk.
Polednak 1990.	• high median income in black tracts	All sites.	For prostate cancer, mortality rate lower for high SES county blacks than nationally. For prostate cancer, blacks have higher risk.
McWhorter et al. 1989.	• median family income • %t below poverty • years of education	14 Cancer sites including prostate	For prostate cancer, poverty failed to explain racial difference. Site of cancer important in SES and race incidence differences. For prostate cancer, blacks have higher risk.
Yu et al. 1989.	• Education • Occupation category	Prostate	No SES-incidence linkage found for blacks. For whites, higher SES increased risk.
Dayal et al. 1985.	• % high school grads • % college grads (>= 25 years of age)	Prostate	Racial difference persisted with adjustments in age, stage and grade, but no increased risk found between races when SES was controlled.
Dayal and Chiu 1982.	• education • income • housing information	Prostate	Racial difference persisted with adjustments in age, stage and grade, but no increased risk found between races when SES was controlled.
Ross et al. 1979.	• educational attainment • category of employment	Prostate and Testicle	No social class differences found in mortality. Blacks have twice incidence of whites.
Earnster et al. 1978.	• >= 25 years of age with some college education	Prostate only	Blacks have higher rates than whites, but racial differences can be described by differences in SES.

Source: Dale, W., S. Vijayakumar, E. F. Lawlor, and K. Merrell. 1996. Prostate cancer, race, and socioeconomic status: Inadequate adjustment for social factors in assessing racial differences. *The Prostate* 29, no. 3: 271-81.

racial-ethnic risk factor interacts with SES in determining incidence rates of prostate cancer.

A comprehensive review by Dale et al. (1996) of 176 socioeconomic status-based studies of prostate and other cancers concluded that there is unclear evidence of the impact of SES on prostate cancer incidence rates. Past studies have shown positive associations, negative associations, and no association whatsoever between SES and prostate cancer (Liu et al. 2001). SES may be linked to the racial and ethnic disparity in incidence rates. Socioeconomic status and class have been shown to be significant predictors of poor outcomes, and blacks are disproportionately represented among the poor and disadvantaged. All in all, the jury is still out on whether SES is a significant risk factor for prostate cancer.

Rural Place of Residence as a Risk Factor for Prostate Cancer

Studies investigating the linkage between rural place of residence and prostate cancer are rare in the literature at the national level and non-existent in Texas as well. A study led by Blair (1985) found that farmers had elevated rates of cancer. There are several reasons why rural place of residence might be a risk factor for prostate cancer, including lower education levels, occupational lifestyle, and exposure to chemical agents including pesticides (Blair 1985). The limited access that those living in the country would have to screening and healthcare would increase mortality rates, but the effects of limited access to healthcare on incidence rates is not yet known. As stated earlier, prostate cancer is most successfully treated in the earlier stages. Earlier treatment would

lead to higher cancer survival rates, but would mot completely explain higher incidence rates in rural populations.

CHAPTER 3

METHODS OF ANALYSIS USED IN THE RESEARCH

The research will employ tested methods of spatial and statistical analysis to answer the following questions: (1) Are there any statistically significant spatial clusters of prostate cancer incidence in Texas at the census tract level? (2) is there any statistically significant association between prostate cancer incidence clusters and socioeconomic status at the census tract level in Texas? (3) is there any significant association between prostate cancer incidence clusters and rural place of residence in Texas?, and (4) is socioeconomic status a significant risk factor for prostate cancer incidence at the census tract level in Texas?

The methods used to investigate these research questions include a spatial scan statistic test for clustering, logistic regression and Poisson regression analysis.

The Spatial Scan Statistic for Cancer Cluster Analysis

The first type of spatial analysis performed on the data was the cluster analysis of prostate cancer incidence. A cluster in this study is defined as a collection of adjacent area units where prostate cancer incidence rates are excessive compared to the rest of the study area (Texas) and the excessiveness is significant (Zhan 2002). An appropriate

approach for this detection of clusters is the spatial scan statistic, developed by Kulldorff (1997, 1998).

The spatial scan statistic was selected for several reasons: (1) the statistic eliminates the problem of pre-selection bias by searching for clusters without specifying their sizes or locations, (2) under a situation where the null hypothesis is rejected, i.e., when the null hypothesis is that no statistically significant spatial clusters exist when cases are assumed to follow a Poisson distribution in space, the approximate location of the cluster that causes the rejection can still be located, (3) it is suitable for inhomogeneous population density (i.e., the state population of Texas), (4) secondary clusters can also be reported, and (5) the method avoids the problem of multiple testing present in most methods, by evaluating the statistical significance of only the most likely cluster and secondary clusters, not every cluster (Kulldorff 1997, Zhan 2001, Zhou 2000). This method is a powerful and proven method in spatial epidemiology (Kulldorff 1997, Kulldorff 1998, Zhan 2001, Zhou 2000). The scan statistic has been used in similar spatial epidemiology studies such as childhood leukemia in Sweden (Hjalmars et al. 1996) and upstate New York (Kulldorff and Nagarwalla 1995), breast cancer in the United States (Kulldorff et al. 1997), childhood cancer in New Mexico (Zhan 2001), and various other cancers in Texas (Zhan 2002, Zhou 2000).

During the execution of the cluster analysis software, the spatial scan statistic imposes a circular window on the map at specified locations in the study area. The window is in turn centered on each of several possible centroids positioned throughout the study region. For each centroid, the radius of the window varies continuously in size from zero to some upper limit, set by the user (Kulldorff 1998). The method calculates

the number of cases expected within the circle based on the at-risk population in the area and the covariates used in the analysis (Zhan 2001). Finally, the most likely cluster is determined through the computation of maximum likelihood ratios (Kulldorff 1998). The likelihood ratio is defined in figure 3.1.

FIGURE 3.1

The Maximum Likelihood Ratio

$$\frac{L(Z)}{L_o} = \frac{\left[\frac{(n_z)}{\mu(Z)}\right]^{n_z}\left[\frac{N-nz}{N-\mu(Z)}\right]^{N-n_z}}{\left[\frac{N}{\mu(A)}\right]^{N}}$$

if $n_z > \mu(Z)$, otherwise $1 / L_o$

Where n_z is the observed number of cases and $\mu(Z)$ is the expected number of cases in cylinder Z. The observed (N) and the expected [$\mu(A)$] number of cases are calculated over the entire study area, across all time periods.

Source: Jacquez, Geoff, and Leah Estberg. 2001. *ClusterSeer: Software for identifying disease clusters, a user guide.* Ann Arbor, MI: TerraSeer.

If a circle has the greatest maximum likelihood ratio and the number of observed cases is more than expected, then the area covered by this circle is considered to be the most likely cluster (Zhan 2001).

SaTScan, a public-domain software package available from the National Cancer Institute (NCI), implements the Kulldorff spatial scan statistic method and is the software used in the cluster analysis portion of this research. SaTScan requires three files for input: a cancer case file, a population file, and a location file. The cancer case file and population file were adapted from the tables used to make the descriptive maps. The underlying population at risk is all males, cross-tabulated by age and race. The location file used is a county or census tract centroid data file with coordinates in geographic decimal degrees. The SaTScan program allows you to run the test for clusters of varying sizes. Before running the analysis, the user must specify the size of the scanning window. This window represents the maximum size of a circle that the test will analyze by percent of the total population in the study area. For this research, window sizes of 5%, 10%, 15% and 50% will be used. SaTScan creates two output files containing the results of the cluster. The first output file is a descriptive text file containing primary cluster locations, secondary cluster locations, P-values and relative rates of risk for the areas in the clusters. SaTScan defines relative rates of risk as the ratio of the number of observed cases divided by the number of expected cases in an area (Kulldorff 1998). The second output file is a text file to be used in a geographic information system for display and further spatial analysis.

Logistic Regression Analysis

The second and third research questions were answered by employing logistic regression analyses. Logistic regression is similar to linear regression in many respects. The primary difference is that in logistic regression the response variable is dichotomous, where a linear regression assumes a continuous variable (McNeil 1996). Logistic regression is an appropriate method for this research because the response variable is a binary variable that represents the status of a census tract as either a tract with a cluster or a tract without a cluster. Another advantage of using logistic regression as a method is that it can handle continuous independent variables, not only dichotomous ones. In this research we will test both dichotomous and continuous exposure variables. Logistic regression is useful for situations in which you want to be able to predict the presence or absence of a characteristic or outcome based on values of a set of predictor variables (SPSS 2000). In this study, logistic regression coefficients were used to estimate odds ratios for each of the independent variables in the model. The dependent variable in each of the logistic regression models was the census tract's status as being a cluster or a non-cluster. The basic logistic regression equation is given in Figure 3.2.

FIGURE 3.2

Logistic Regression Equation

$$\text{log odds} = a + \sum_{i=1}^{k} b_i x_i$$

where the log odds, is required to be a linear function

of the risk factor magnitude for cases 1 through k

Source: McNeil, Don. 1996. *Epidemiological research methods.* New York: John Wiley & Sons.

Poisson Regression Analysis

Poisson analysis was utilized in this study to ascertain any correlations between socioeconomic status and prostate cancer incidence rates or prostate cancer clusters at the county and census tract level. The concept was developed by the French mathematician C.D. Poisson (1837) in the 19th century. The basic idea of Poisson regression was outlined by Coleman (1964). Poisson is used to estimate the number of expected occurrences in an area, which is necessary in this study to calculate standardized incidence ratios, as discussed in section 1.4.1.1. Poisson is a skewed distribution appropriate and useful for phenomena that have a very small probability of occurring on any particular trial, but for which an extremely large number of trials are available (while the product of the two numbers is moderate) (SPSS 2001). Poisson is also useful when the populations at risk differ for each of the covariate patterns. More succinctly, when population distributions differ by race, ethnicity, or age, from place to places, as they do in this study, Poisson analysis is recommended (Egret 1999). The small probability of occurrence of prostate cancer in the Texas population and our large number of trials makes this method a logical choice.

CHAPTER 4

GEOGRAPHIC DISTRIBUTION OF PROSTATE CANCER INCIDENCE CLUSTERS IN TEXAS

Introduction

The analysis in this chapter aims to meet the objective of obtaining a clear view of the significant clusters of overall prostate cancer incidence at the census tract level in Texas. The remainder of this chapter is organized into four sections. Section two describes the source data used in this part of the study. Section three discusses the method used, data preparation, and analytical procedure. Section four describes the results and section five discusses the conclusions.

Source Data

Three data sets were needed for the spatial scan statistic cluster test. These data sets include cancer incidence data, population data, and geographic data.

Incidence Data

Cancer incidence data were obtained from the Texas Cancer Registry in the Texas Department of Health for the period 1990-1997. Prostate cancer incidence data were extracted from the original data set based on the ICD-9 (International Classification of

Diseases 9th Revision) codes for prostate cancer. The source data from the Texas Department of Health contained a total of 45 attributes. The attributes used in this chapter include case, race, ethnicity, age at diagnosis, street address, city, state, and zip code.

Population Data

The population data for each census tract were obtained from the U.S. Bureau of the Census web site (Census 2001). There were a total of four age groups (18-24, 25-44, 45-64, 65+) and three race-ethnic categories (white, black, and Hispanic) used in this research. The 1990 population data for each of the twelve age-race-ethnicity categories were downloaded for each of the 4,045 census tracts in Texas. The race 'white' is assumed to be white non-Hispanics in this study and the race 'black' is considered to be black non-Hispanics.

Geographic Data

The goegraphic data used in the cluster analysis consists of the 1990 shape-file of the 4,045 Texas census tracts and a tract centroid file, both obtained from the Environmental Systems Research Institute (ESRI) Data & Maps Media Kit CD-Rom. The centroid file was generated using an ArcView Avenue script and contains approximations of the centers of indiviudal tracts represented by lattitude and longitude in decimal degrees. The centroid file also contains an attribute for people per square mile in each census tract, which was used for rural and urban coding.

Procedure of Analysis

Data Preparation

The SaTScan program requires that three separate files be constructed for each cluster test in each category.

For the centroid file, no further preparation was necessary and an identical centroid file was used in all cluster analyses.

The case file required some preparation. The prostate cancer incidence source data from the Texas Department of Health did not contain census tract-level information for each case, of which there were 57,529 in the years 1990-1997. Of these, 44,725 were white non-Hispanic, 6,267 were black, and 5,679 were Hispanic. When other races and incorrect age data were removed, there were 56,341 total cases used in the study. In order to perform the analysis, the 56, 341 cases of prosate cancer over the eight year period, 1990-1997, were geocoded. Geocoding was performed using ArcView GIS version 3.2 and the ArcView StreetMap geocoding extension, both products of ESRI corporation. Of the 56,341cases, 41,528 (74%) were sucessfully geocoded and matched to census tracts. Many of the addresses were incomplete or only contained data that could not be accurately geocoded (post office boxes, inaccurate street names, etc.). Once the cases were geocoded, they were linked to specific census tracts by a spatial join to a census tract file. All cases were sucessfully attributed to census tracts.

The age and race portions of the case file also required some preparation. Ages were re-coded into four separate categories (18-24=1, 25-44=2, 45-64=3, and 65+=4) and races were re-coded into three categories (white=1, black=2, hispanic=3).

The resulting case file for the cluster analysis contained the following information: census tract id, number of cases (always 1 per line), year of diagnosis, age covariate (1, 2, 3 or 4), and race covariate (1,2 or 3). The case file contained 41,528 records; one for every case.

The population file was based on data from the U.S. Census Bureau. First, population data were extracted from the 4,045 tables of Texas census tracts from 1990. The data extracted included populations for each of the twelve age-race-ethnicity (all male) groups in each tract. To estimate the population for the tracts for the years 1991-1997, the total male population growth rate for Texas was determined and that linear growth rate was applied to the 1990 numbers. Using Microsoft Access version 9.0, a population file was built containing one line for each year (8 years) , race-sex ethnic category (12 categories) and each census tract (4,045 tracts). The resulting population input file contained 388,320 records.

Analytical Procedure

One case file and one population file was prepared for the SaTScan program. The SaTScan software analysis was performed using circles of varying size, starting with 5% of the population at risk as the maximum circle size, and increasing the circle size in increments of 5% until the resulting clusters remained the same when the maximum circle sizes increased. In the case where a 5% maximum circle size is selected, SaTScan only draws cluster circles up to 5% of the total population, but uses the cases and populations of each age-race-ethnicity for the total area of Texas to generate expected

incidence rates. This technique was used by Zhou (2000) and ensures that the maximum-sized cluster is detected.

SaTScan also allows for the user to choose several other options when analyzing the data. In every run of the analysis, a purely spatial analysis was conducted, using the Poisson probability model. The area was scanned for high rates of cancer. The time period of the study specified was January 1, 1990 to December 31, 1997. The number of Monte Carlo replications was set to 999 in each run. A Pentium II PC with a 400 MHz processor running the Windows 2000 operating system was used to conduct the analysis. The running times varied from six minutes thirty-eight seconds to twelve minutes and twelve seconds.

The two output files consist of a file for use in a geographic information system and a file with details of each cluster. These files will be used in the further analysis.

Results

The analysis resulted in a significant primary cluster and several secondary clusters. A cluster is significant when it has a p value of less that 0.05. SaTScan assigns each cluster a cluster ID number. An ID number of 1 indicates a primary cluster and subsequent ID numbers indicate secondary clusters. These primary and secondary clusters are summarized in Table 4.1 and detailed in Table 4.2.

TABLE 4.1

Summary of Cluster Analysis Data

Area Investigated	Texas
Level Investigated	Census Tracts
Type of Analysis	Purely Spatial
Probability Model	Poisson
Scan for Areas With	High Rates
Start and End Date	1/1/1990 to 12/31/97
Number of Monte Carlo Replications	999
Number of Census Tracts	4,045
Annual Population	7, 042, 100
Total Cases	41, 528
Annual Incidence per 100,000	73.7
Maximum Spatial Cluster Size	10.00

TABLE 4.2

Significant Clusters of Prostate Cancer Incidence at the Census Tract Level in Texas, 1990-1997

Cluster ID	Cluster Information [center lat long/radius (km)/# tracts]	Pop.	Observed Cases	Expected Cases	Relative Risk	Annual incidence per 100,000	Log Likelihood Ratio	p value	Significance level (%)
1	(29.543 N, 98.471 W)/ 7.55 / 32	59608	961	298.91	3.215	237.0	465.54	0.001	.1
2	(29.629 N, 95.635 W) / 25.60/ 210	430895	3121	1829.62	1.706	125.7	396.61	0.001	.1
3	(33.027 N, 96.976 W) / 29.25 / 277	542258	3240	2096.23	1.546	113.9	283.78	0.001	.1
4	(30.499 N, 97.973 W) / 41.18 / 177	282433	1945	1151.98	1.688	124.5	233.56	0.001	.1
5	(31.528 N, 97.209 W) / 2.20 / 3	3635	129	23.89	5.400	398.0	112.56	0.001	.1
6	(31.594 N, 97.131 W) / 0.00 / 1	401	44	1.78	24.767	1825.7	99.017	0.001	.1
7	(29.831 N, 94.635 W) / 63.42 / 268	437378	2897	2393.91	1.210	89.2	52.753	0.001	.1
8	(31.679 N, 99.135 W) / 14.36 / 5	1195	46	7.06	6.519	480.5	47.309	0.001	.1
9	(33.540 N, 101.913 W) / 6.20 / 31	51657	459	281.48	1.631	120.2	47.30	0.001	.1
10	(33.482 N, 94.295 W) / 22.97 / 14	17524	188	84.59	2.222	163.8	46.855	0.001	.1
11	(31.568 N, 97.168 W) / 0.00 / 1	1295	19	0.63	30.089	2217.9	46.314	0.001	.1
12	(30.501 N, 96.252 W) / 0.00 / 1	883	29	2.66	10.889	802.6	42.915	0.001	.1
13	(27.653 N, 97.196 W) / 21.76 / 22	49828	410	264.20	1.552	114.4	34.630	0.001	.1
14	(32.387 N, 99.677 W) / 10.03 / 17	21687	246	140.61	1.750	129.0	32.344	0.001	.1

TABLE 4.2 CONTINUED

Cluster ID	Cluster Information [center lat long/radius (km)/# tracts]	Pop.	Observed Cases	Expected Cases	Relative Risk	Annual incidence per 100,000	Log Likelihood Ratio	p value	Significance level (%)
15	(31.115 N, 97.342 W) / 0.00 / 1	4195	17	1.11	15.370	1132.9	30.559	0.001	.1
16	(31.896 N, 106.246 W) / 17.64 / 26	88034	513	357.46	1.435	105.8	30.076	0.001	.1
17	(35.154 N, 101.911 W) / 4.01 / 16	18965	215	124.70	1.724	127.1	26.916	0.001	.1
18	(30.655 N, 96.331 W) / 0.00 / 1	516	27	4.43	6.094	449.2	26.235	0.001	.1
19	(31.881 N, 102.358 W) / 4.13 / 14	22703	241	145.86	1.652	121.8	25.987	0.001	.1
20	(31.827 N, 106.578 W) / 10.24/ 14	40926	341	235.72	1.447	106.6	20.763	0.001	.1
21	(28.875 N, 97.023 W) / 9.63 / 11	19749	194	127.37	1.523	112.3	15.051	0.002	.2
22	(31.429 N, 100.492 W) / 6.30 / 12	19623	194	127.52	1.521	112.1	14.971	0.002	.2
23	(32.254 N, 95.362 W) / 10.49 / 14	20854	238	166.62	1.428	105.3	13.542	0.010	1
24	(34.187 N, 101.729 W) / 2.22 / 3	5360	76	39.88	1.906	140.5	12.901	0.013	1.3

The analysis revealed the primary cluster of prostate cancer incidence centered in northern Bexar County in San Antonio, Texas. The cluster contained 32 census tracts with 3.2 times the expected number of cases present in the cluster. The cluster significance was 0.1% (p=0.001).

Further, the analysis revealed 23 significant secondary clusters in other regions of Texas. The relative risk rates in these secondary clusters ranged from 1.4 to 30.2 and the size of the clusters ranged from only one tract to 279 tracts. The primary and secondary clusters are displayed in Figures 4.1, 4.2, and 4.3.

FIGURE 4.1

Texas Counties Containing Clusters of Prostate Cancer Incidence, 1990-1997

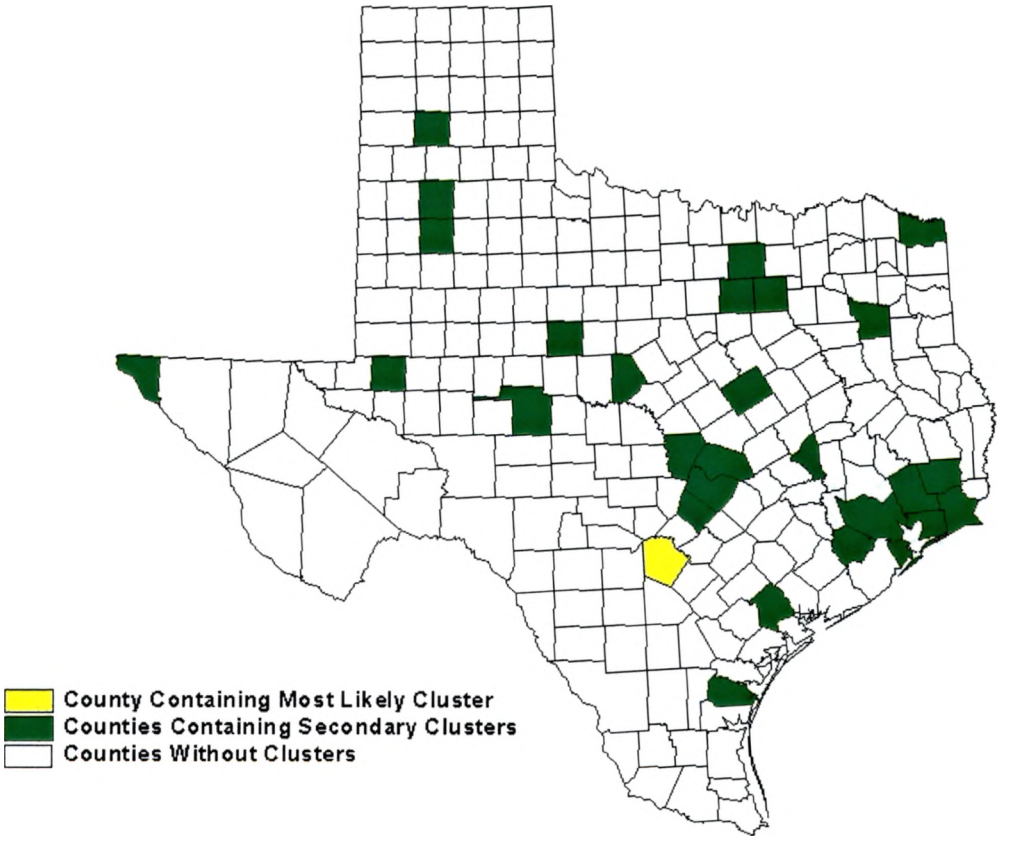


FIGURE 4.2

Most Likely and Secondary Clusters of Prostate Cancer in Texas
at the Census Tract Level, 1990-1997

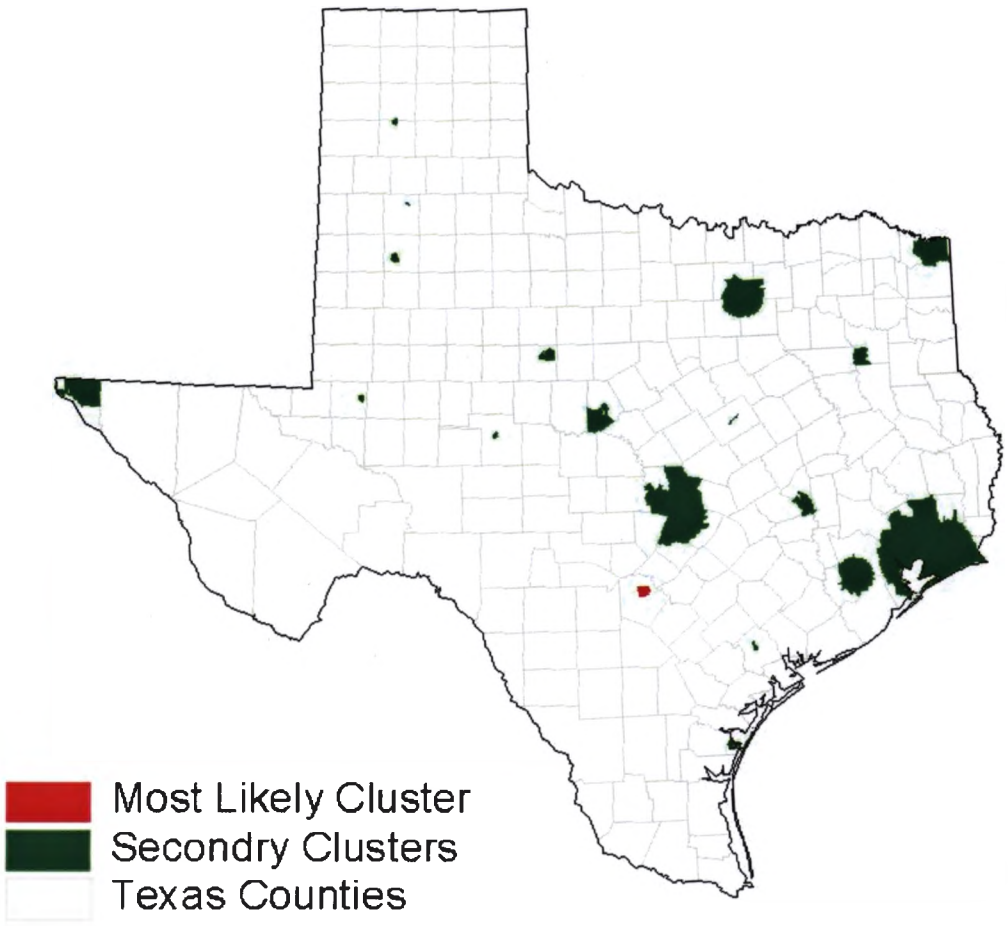
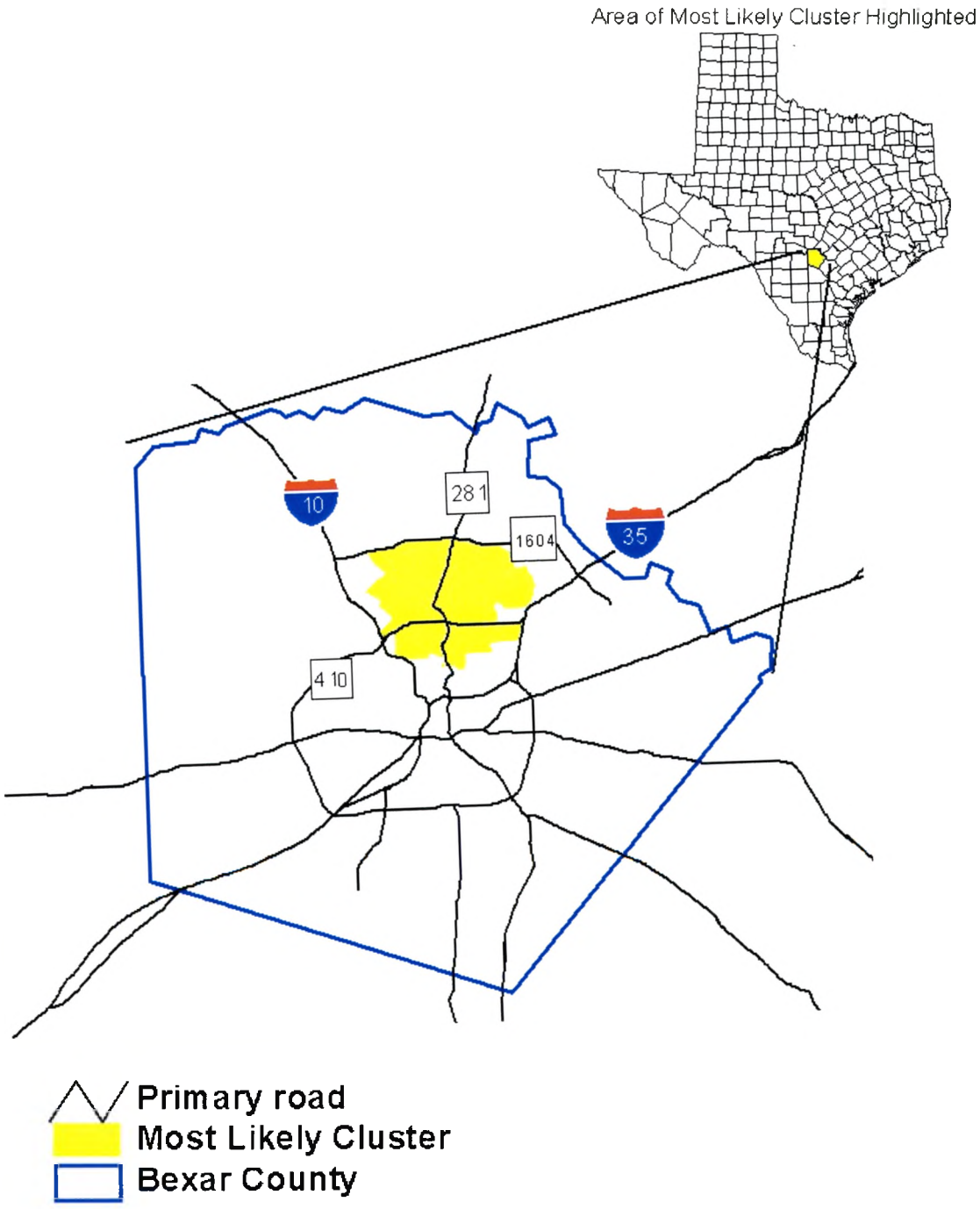


FIGURE 4.3

Most Likely Clusters of Prostate Cancer in Texas at the Census Tract Level, 1990-1997



Conclusions and Discussion

Population data from the U.S. Census Bureau and prostate cancer incidence data from the Texas Cancer Registry at the Texas Department of Health were used to compute a spatial scan statistic adjusted for age, race, and ethnicity for the years 1990-1997. Results of these analyses indicated that 24 significant clusters of prostate cancer incidence were detected at the census tract level. These clusters consisted of one primary cluster and 23 secondary clusters. The primary cluster contains 32 census tracts and the secondary clusters contained 1, 171 census tracts.

The most likely (primary) cluster was located in northern Bexar County (northern San Antonio) and the secondary clusters were located in various regions, mostly urban, around the state. Figures 4.1, 4.2 and 4.3 show the spatial distribution of significant clusters. No significant clusters were detected in southwestern Texas or along the border region, except for El Paso.

In the future, it might be useful to consider different age-race-ethnicities in searching for any clusters among specific age groups or race-ethnic groups. Zhou (2000) found that lung cancer mortalities cluster in certain age-race groups correlated well to certain industrial occupations, like forestry. This research on prostate cancer incidence adjusts for all of the age groups and race-ethnicities, but does not draw out and analyze each one. Examining this future direction might provide some additional insight into the nature of the clusters discovered in this thesis.

Second, another cluster analysis should be performed when complete census data and case data is available for all census tracts in Texas. This study was limited in that 1990 census data were used as a baseline to estimate the growth of the male population in

Texas. In addition, only 1990-1997 prostate cancer incidence data were available at the time of the study. A complete set of data for 1990-2000 might yield more accurate results.

A third course that future research might take would be to investigate the *lack* of clusters along the border region, except for El Paso. As noted in Chapter two, Hispanics generally have lower rates of prostate cancer incidence than white and black men, but this study was adjusted for race and ethnicity. The 2000 U.S. Census data for the border region will be updated with more accurate figures. This newly available data might be applied to a study of prostate cancer along the Texas-Mexico border.

Often times, cancer clusters are related to risk factors associated with a particular area. The next chapter will investigate environmental risk factors such as socioeconomic status and rural place of living and their association with prostate cancer incidence.

CHAPTER 5

RELATIONSHIPS BETWEEN PROSTATE CANCER INCIDENCE CLUSTERS, PROSTATE CANCER INCIDENCE RATES, SOCIOECONOMIC STATUS, AND RURAL PLACE OF RESIDENCE

Introduction

Socioeconomic conditions and rural place of living have been associated with cancer rates in various degrees of significance and intensity (Dale et al. 1996).

The objective of this chapter is to address the second, third and fourth research questions dealing with statistically significant association between prostate cancer incidence clusters and socioeconomic status at the census tract level in Texas, significant associations between prostate cancer incidence clusters and rural place of residence in Texas, and socioeconomic status as a significant risk factor for prostate cancer incidence at the census tract level in Texas.

This chapter contains two separate statistical analyses. The first analysis is used to determine relationships between *clusters* of prostate cancer incidence, socioeconomic status and rural place of residence. The second analysis is used to determine relationships between the incidence rates in each census tract for each age-race-ethnic category and socioeconomic status.

The remainder of this chapter is organized as follows. Data sets used in this portion of the research are introduced in section two. Section three discusses the data

preparation. Section four reviews the statistical analyses being conducted. Section five describes the results and section six draws conclusions based on the results.

Source Data

Four types of source data were needed in this section of the research. The types of data included: incidence data, population data, geographic data, and socioeconomic data.

Incidence Data

The incidence data, or case data, was interpreted from the case data used in the cluster analysis. A total of 41,528 cases from 1990-1997 for the 4,045 census tracts in Texas were used in this section of the analysis.

Population Data

Population data were obtained at the census tract level. A derivation of the data used in the cluster analysis was used. Populations for each census tract in each age-race-ethnicity category were extrapolated from the cluster analysis data.

Geographic Data

One risk factor that was examined in this study was rural versus urban living location. Although the reasons are still unknown, some studies have shown elevated rates of prostate cancer in rural areas and among farmers (Blair, Malker, and Cantor 1985; Carter, Carter, and Isaacs 1990; Hayes 2001). The definition of what constitutes a rural

area is a vague concept and varies from one perspective to the next. For this study, the 2000 U.S. Census Bureau definition of urban were used to make the distinction.

> For Census 2000, the Census Bureau classifies "urban" as all territory, population, and housing units located within an urbanized area (UA) or an urban cluster (UC). It delineates UA and UC boundaries to encompass densely settled territory, which consists of: (1) core census block groups or blocks that have a population density of at least 1,000 people per square mile and (2) surrounding census blocks that have an overall density of at least 500 people per square mile. "Rural" consists of all territory, population, and housing units located outside of UAs and UCs. It contains both place and nonplace territory. Geographic entities, such as census tracts, counties, metropolitan areas, and the area outside metropolitan areas, often contain both urban and rural territory, population, and housing units (Census 2001).

More succintly, in the terms of this research, an urban census tract was one that contained 1,000 or more people per square mile. A rural tract was defined as one with less than 1,000 people per square mile. Geographic data of population per square mile is from the U.S. Census Bureau and is interpreted from the geographic centroid file used in the cluster analysis.

Socioeconomic Status Data

Socioeconomic status (SES) data were obtained from the 1990 and 2000 U.S. Census and estimates by the Texas State Data Center. Based on previous studies in the literature (Baquet et al. 1991; Dale et al. 1996; Dayal and Chiu 1982; Dayal, Polissar, and Dahlberg 1985; Earnster et al. 1978; Hoffman et al. 2001; McWhorter et al. 1989; Polednak 1990; Ross et al. 1979; Yu et al. 1988), SES was calculated using two census variables: educational attainment and median household income. These two factors are sufficient predictors of SES and were obtained from the U.S. Census Bureau (Baquet et

al. 1991; Dale et al. 1996; Dayal and Chiu 1982; Dayal, Polissar, and Dahlberg 1985; Earnster et al. 1978; Hoffman et al. 2001; McWhorter et al. 1989; Polednak 1990; Ross et al. 1979; U.S. Census Bureau 2001; Yu et al. 1988).

Education as a Socioeconomic Factor

The education socioeconomic factor variable required preparation. The census collects data on educational attainment at differing age levels. The data set chosen for education in this study are those age 25 and over with some college education. For comparison purposes, the raw education number furnished by the census is converted to a percentage by dividing the raw number of people age 25 and over with some education by the total population age 25 and over living in that census area. The education levels were broken into quartiles at the state level.

Income as a Socioeconomic Factor

The other socioeconomic indicator variable was average household income. No conversion of the raw data for this variable was required. The U.S. Census Bureau furnishes this variable that represents the median value of the household income in each census tract. Median household income values were also broken into quartiles based on the statewide distribution.

Data Preparation

Two separate sets of data were prepared for the two phases of the statistical analysis. The first phase evaluated the census tracts that contained clusters found in the

cluster analysis using logistic regression. For this first phase, a cluster recode variable (the dependent variable), a geographic urban-rural variable, and two socioeconomic variables were prepared. The second phases of the statistical analysis evaluated all census tracts separately by age-race-ethnicity using a Poisson analysis. For this second phase, a case variable, a population variable, and two socioeconomic variables were prepared. Table 5.1 summarizes the recoding of the variables.

Preparation of the Dependent Variable for Cluster Logistic Regression

For the cluster logistic regression, the dependent variable was a derivative of the cluster variables generated in the analysis found in Chapter four. The cluster variable was recoded as follows. If a particular census tract contained a primary (most likely) or a secondary cluster, then the tract was recoded as a one. If a tract contained no clusters, the tract was recoded as a two. The dependent variable was recoded for ease of interpretation of the results. Tracts with clusters are displayed in Figure 5.1.

Case Data Preparation for Incidence Poisson Analysis

For the Poisson analysis, the cases for each of the 4,045 census tracts were divided into the twelve age-race-ethnicity categories for the years 1990-1997. The cases for each year were summed. The resulting file contains 48, 540 records with covariate columns for age and race-ethnicity.

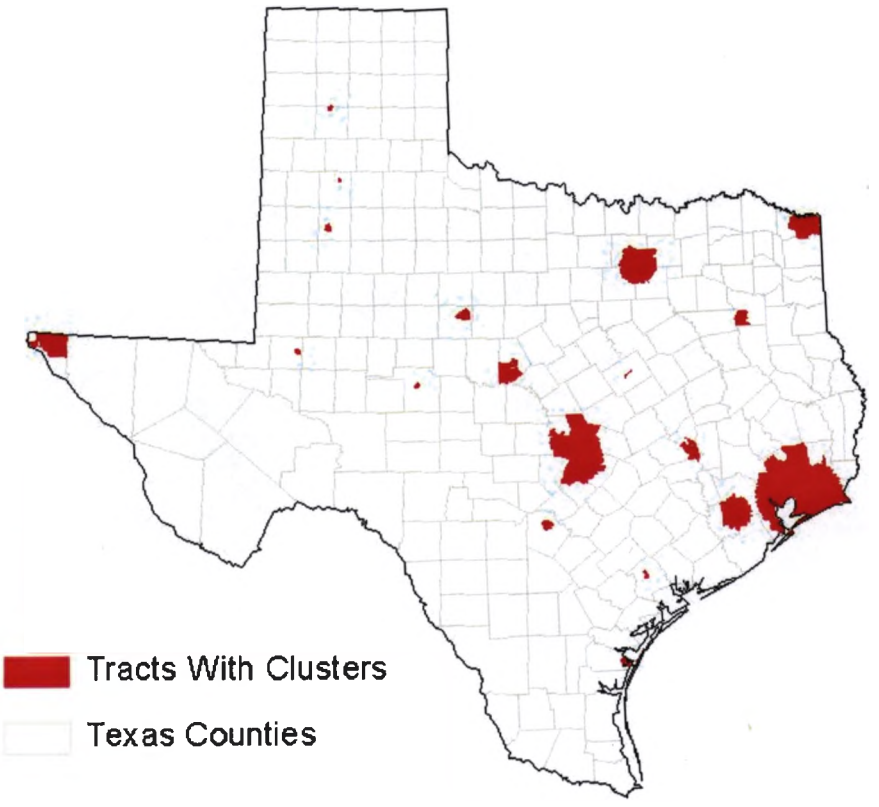
TABLE 5.1

Recoded Variables Used in Statistical Analyses

Variable (Variable Name)	Dependent / Independent	Description	Original Value(s)	Recoded Value
Cluster	Dependent	Tracts With Clusters	Contained Clusters	1
	Dependent	Tracts Containing No Clusters	No Clusters	2
Cases	Dependent	Total Cases in Tract	Number of Cases in Tract, 1990-97	N/A
People Per Square Mile	Independent	Urban Census Tracts	$\geq \frac{1{,}000}{mi^2}$ Urban	1
	Independent	Rural Census Tracts	$< \frac{1{,}000}{mi^2}$ Rural	2
Median Household Income	Independent	Highest Quartile (Most Income)	≥ $33,114	1
	Independent	Medium – High	$24,448 - $33,113	2
	Independent	Medium - Low	$18,277 - $24, 447	3
	Independent	Lowest Quartile (Least Income)	≤ $18, 276	4
Educational Attainment	Independent	Highest Quartile (Most Educated)	≥ 61%	1
	Independent	Medium – High	39% - 60%	2
	Independent	Medium - Low	26% - 38%	3
	Independent	Lowest Quartile (Least Educated)	≤ 25%	4

FIGURE 5.1

Census Tracts Containing Clusters of Prostate Cancer Incidence, 1990-1997



Population Data for Incidence Poisson Analysis

For the Poisson analysis, a person-years population variable was created. The population variable for the incidence Poisson analysis contains the individual populations of each of the 4,045 census tracts divided into the twelve age-race-ethnicity categories for the total of the years 1990-1997. The resulting variable file contains 48, 540 records with covariate columns for age and race-ethnicity co-variates.

Geographic Data for Cluster Logistic Regression

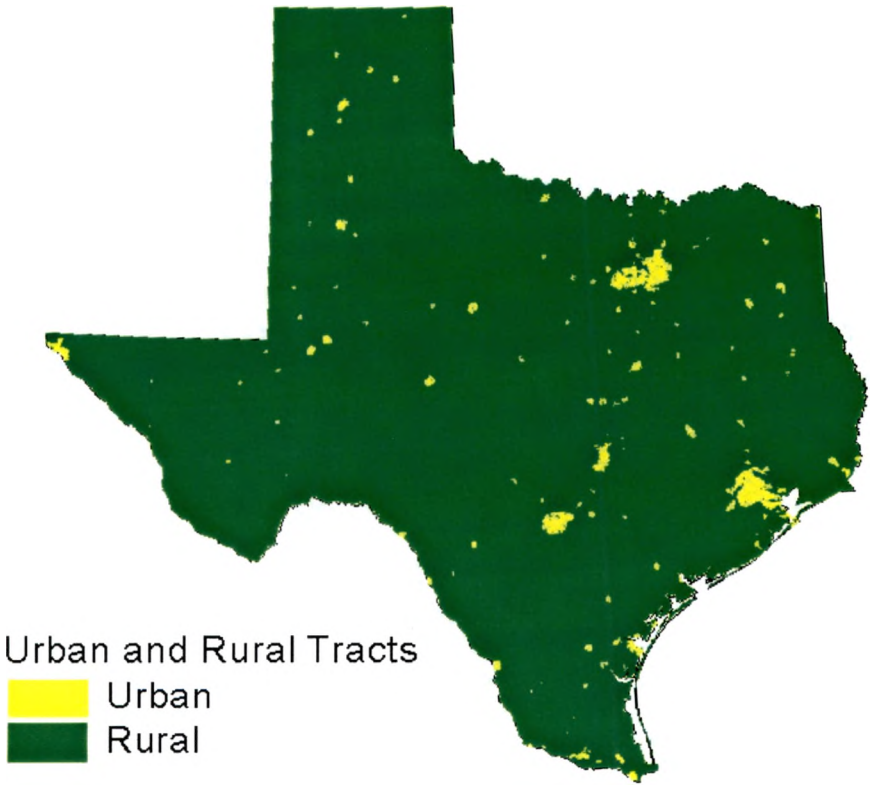
The designation of urban or rural place of residence is based on the U.S. Census definition as described earlier in this chapter. All tracts that have a population of 1,000 or more persons per square mile were designated as urban. All other tracts were designated as rural. The resulting variable comprised of 4,045 records with each tract coded as urban or rural in the following fashion. If a tract contained 1,000 persons or more per square mile, then the tract variable was recoded as a one, for urban. If a tract contained less than 1,000 persons per square mile, then the tract was recoded as a two, for rural. The geographic independent variable was recoded for ease of interpretation of results. Rural and urban areas are shown in Figure 5.2.

Socioeconomic Data for Cluster Logistic Regression and Incidence Poisson Regression

Socioeconomic data were made up of two factors, income data and education data. The reason for selecting these two factors was discussed in Chapter 2. All socioeconomic data were gathered from the 1990 U.S. Census Summary tape File 3 (STF 3) sample data.

FIGURE 5.2

Rural and Urban Census Tracts in Texas, 1990



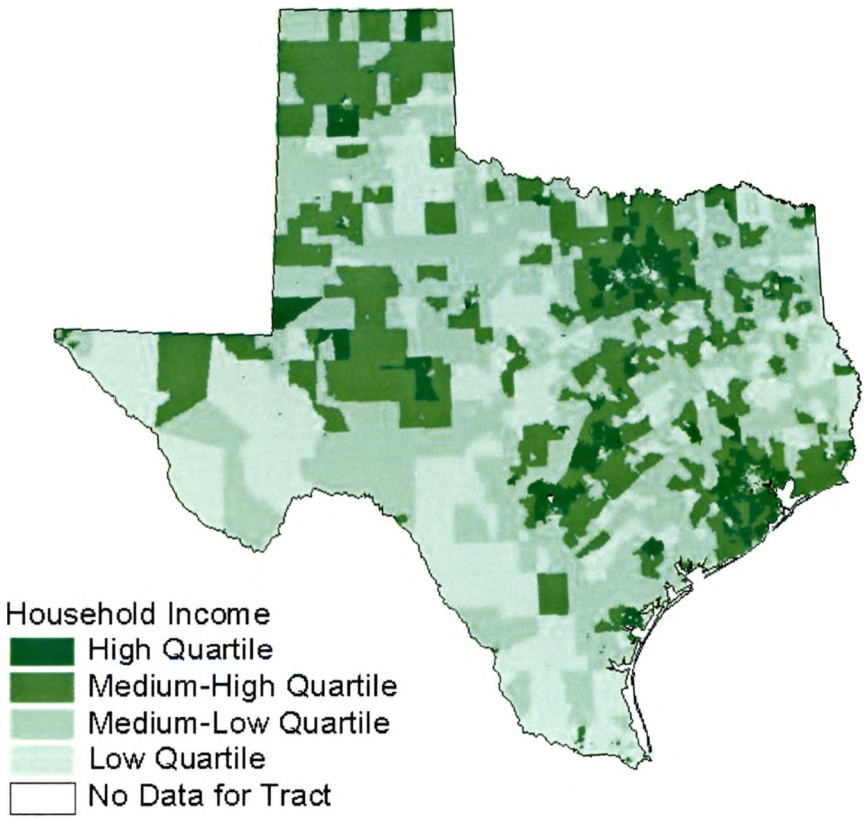
Where urban tracts are defined as those tracts with 1,000 or more people per square mile. Rural tracts are defined as those with less than 1,000 people per square mile.

The U.S. Census category for median household income (Census Table P080A) was used in this research as the base data for the income factor. The variable was broken into quartiles based on the median income distribution of by all Texas census tracts and recoded into the following categories. The highest income quartile were tracts with a median household income of $33,114 and higher. This highest class was recoded as a one. The next highest quartile of classes were tracts with a median household income of between $24,448 and $33,113. This second highest quartile was recoded as a two. Tracts with median household income between $18,277 and $24, 447 were recoded as a three and the lowest income quartile, those with median household incomes below $18, 276 were recoded as a four. Figure 5.3 displays the income quartiles by census tract in Texas.

The other measure used in as an indicator of socioeconomic status was the percentage of the population of the tract, age 25 and over, that received some form of college education. The U.S. Census category for education level (Census Table P057) was used. The variable was broken into quartiles based on the median income distribution of all Texas census tracts and recoded into the following categories. The highest education quartile were those tracts with 61% or more of the population over 25 that received some college education. This class was recoded as a one. The next highest class were tracts with 39% to 60% of the population over 25 years of age that received some college education and were recoded as two. Those tracts with education levels between 26% and 38% were in the third quartile and coded as a three, and those in the lowest quartile, with education populations age 25 and over at levels of 25% and below were recoded as four. Figure 5.4 displays the statewide education levels by quartile.

FIGURE 5.3

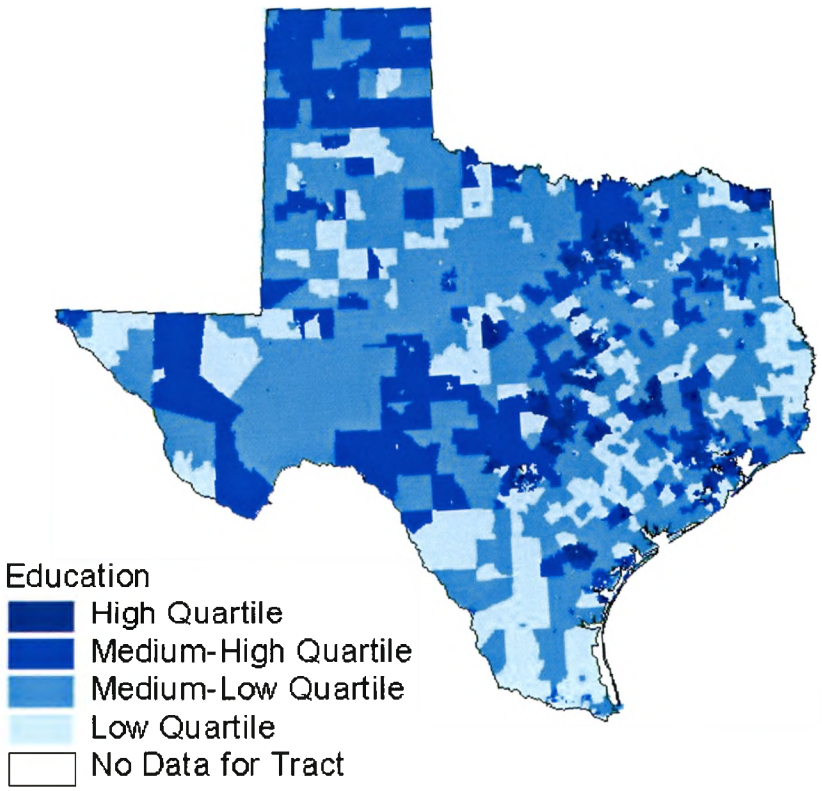
Median Household Income Quartiles by Census Tract in Texas, 1990



Where High > $33,113; Medium High $24,448 - $33,113; Medium Low $18,277-$24, 447; Low < $18, 276

FIGURE 5.4

Education Quartiles by Census Tract in Texas, 1990



Where High > 60%; Medium High 39-60%; Medium Low 26-38%; Low <26%

Statistical Analyses

The first research question of determining the locations of prostate cancer incidence was answered using spatial analysis. The three remaining research questions concern questions of relationships between the clusters of prostate cancer incidence and rural place of residence, relationships between clusters of prostate cancer incidence and socioeconomic status, and the relationship between the rates of prostate cancer incidence in each tract and socioeconomic status as a whole, when adjusted for age-race-ethnicity. These questions must be answered using statistical analysis.

To answer the questions about relationships between socioeconomic conditions, place of residence and clusters, logistic regression was the method employed. The question about the relationship between prostate cancer incidence and socioeconomic status adjusted for age, race, and ethnicity was answered using a Poisson analysis. These two methods were previously outlined in Chapter three.

For both the logistic regression and the Poisson analysis, a Pentium II PC with a 400 MHz processor running the Windows 2000 operating system was used to conduct the analysis. SPSS version 10.0 statistical software was the analysis package used in both analyses.

Logistic Regression

One file was prepared for the logistic regression. The file contained 4,045 records (one per tract). The dependent variable for the regression was the recoded cluster variable. The three independent variables were the urban/rural recoded variable, the household income quartile variable and the education quartile variable.

A binary logistic regression model was selected because of the dichotomous dependent variable. Education quartile and household income quartile variables were defined as categorical variables. The change contrast was set to indicator with the last category (lowest socioeconomic status) selected as the reference category. The confidence interval for the test was 95% and the test employed a Hosmer-Lemeshow goodness of fit test.

Poisson Analysis

The Poisson analysis file contained one record for each of the twelve age-race-ethnicity categories for each census tract, totaling 48,540 records.

A Poisson general log linear analysis was performed using education quartile and household income quartile as the factors in the analysis. The total population in each census tract for each age-race-ethnicity covariate defined the cell structure. The number of cases in each tract for each age-race-ethnicity defined the weight cases for the Poisson regression. Age and race recoded variables were selected as covariates, but were not entered into the equation. The confidence interval was set to 95%.

Results

The results of the statistical analysis are separated into two sections. The first section describes the results of the logistic regression using census tracts that contained clusters of prostate cancer incidence as the dependent variable. The second section outlines the results of the Poisson analysis of each individual census tract by age-race-ethnicity.

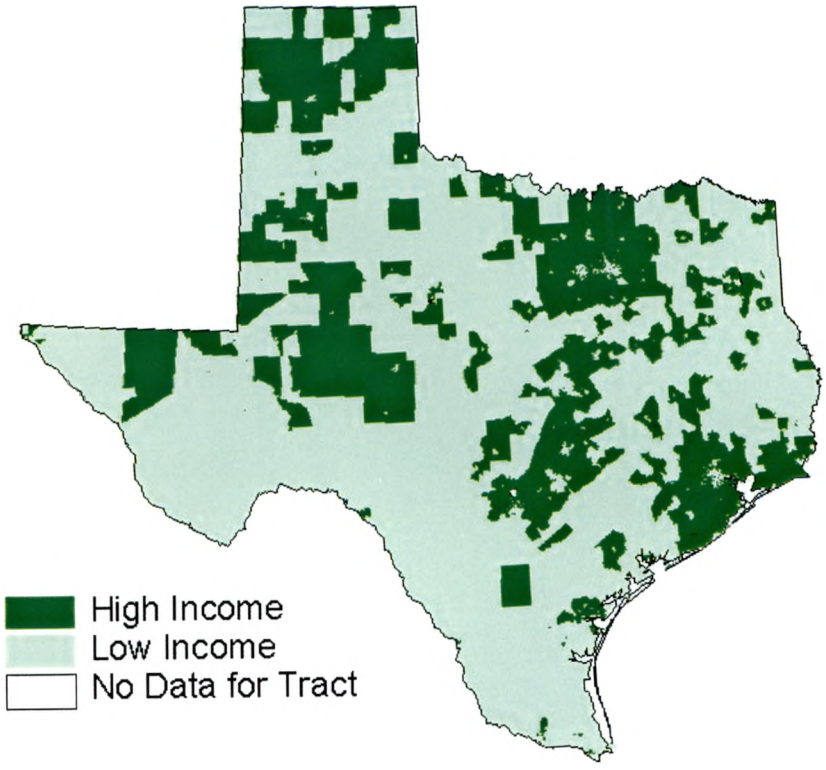
Relationship Between Prostate Cancer Incidence Clusters, Socioeconomic Status, and Rural Place of Residence

Forward binomial logistic regression was performed to determine if household income, education, or rural place of residence were predictors of a census tract containing a cluster of prostate cancer incidence.

The initial run of the logistic regression, utilizing socioeconomic variables broken into quartiles, did not yield an acceptable goodness of fit model (Chi-square = 34.025, *df*=*8,* Sig.=0.000). The literature suggests that one should collapse the independent variables in this situation (Mertler and Vannatta 2002). Data screening led to the reduction of the two socioeconomic factor variables from four categories to two categories each. The four quartiles of household income and the four quartiles of education were collapsed from four categories each to two categories, high and low. Figures 5.5 and 5.6 show the geographic distribution of the compressed variables.

FIGURE 5.5

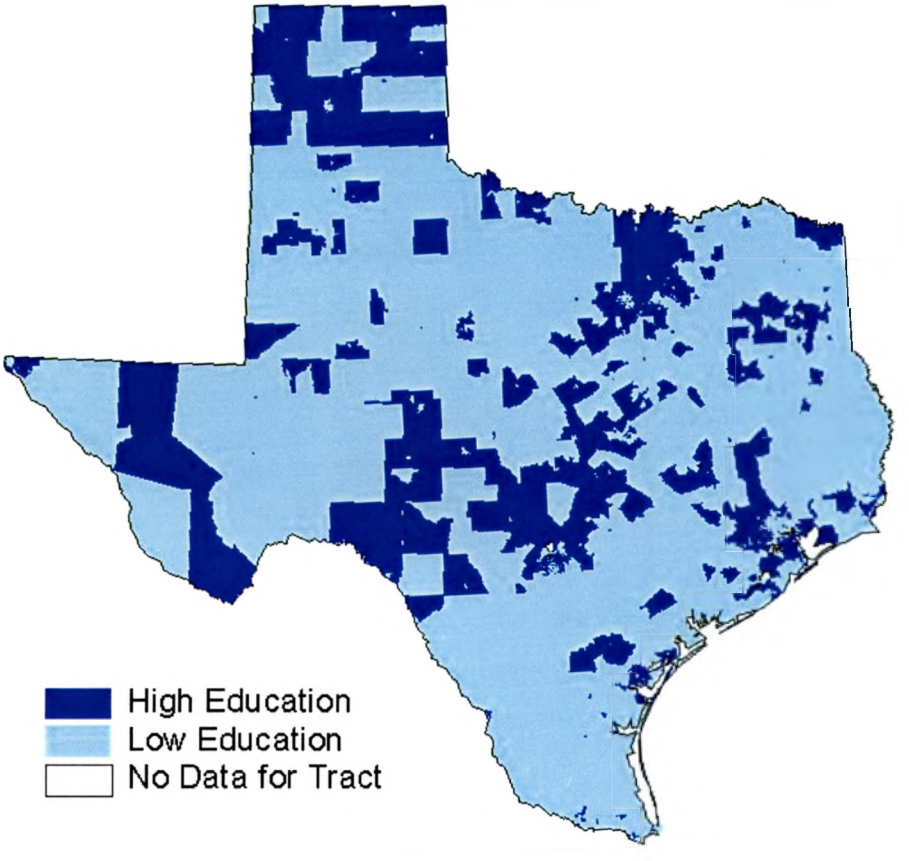
High and Low Income Census Tracts, 1990



Where High >= $24,447; Low < $24, 447

FIGURE 5.6

High and Low Household Income Census Tracts, 1990



Where High > =39%; Low <38%

The resulting regression results using the transformed variables indicated a model with a closer fit, fewer degrees of freedom, and higher significance (Chi-square = 13.306, *df*=5, Sig.=0.021). The model correctly classified 74.1% of the cases. Wald statistics indicate that all variables significantly predict counties with clusters of prostate cancer. However, odds ratios for these independent variables indicated only a small positive change in the likelihood of a county with a cluster of prostate cancer incidence.

Risk estimates of the three independent variables were calculated using cross tabs. All risk estimates were significant. The overall odds ratio for higher education was 5.936 (5.046 – 6.984, 95% C.I.). The overall odds ratio for household income was 3.991 (3.428-4.646, 95% C.I.). The overall odds ratio for urban-rural place of residence was 3.361 (3.108 –4.241, 95% C.I.). Table 5.2 lists all risk estimates with Pearson chi-square values, degrees of freedom, and 95% confidence intervals.

The logistic regression model yielded similar odds ratio results to the univariate results. As predictors of a county having a cluster of prostate cancer incidence, those with higher education level have the highest odds ratio of 3.306 (2.717-4.021, 95% C.I). High income census tracts have an odds ratio of 2.513 (1.785 – 2.598, 95% C.I). Urban place of residence yields an odds ratio of 3.058 (2.585-3.618, 95% C.I). Regression coefficients, Wald statistics, significance, odds ratios and 95% confidence intervals of odds ratios are presented in Table 5.3.

TABLE 5.2

Risk Estimates for Independent Variables

Variable	Risk Estimate	95% C.I.	
		High	Low
High Education	5.936	5.046	6.984
Low Education	1.000	-	-
High Income	3.991	3.428	4.646
Low Income	1.000	-	-
Urban Residence	3.631	3.108	4.241
Rural Residence	1.000	-	-

TABLE 5.3

Odds Ratio Results of Logistic Regression

Variable	B	Wald	p	Odds Ratio	95% C.I.	
					Low	High
High Education	1.196	142.942	0.000	3.306	2.717	4.021
High Income	0.767	64.169	0.000	2.153	1.785	2.598
Urban Residence	1.118	169.914	0.000	2.585	2.585	3.618

The final stage of the logistic regression was to employ interaction variables between the independent variables in the model. With the aim of increasing the fit of the model, interaction variables were introduced to the logistic regression model. The first interaction variable created was a place of living and education variable. The second interaction variable created was a place of living and income variable. The third interaction variable was a variable made up of the interaction between income and education. These variables were entered as independent variables, along with the original three variables: urban or rural place of residence, high or low household income, and high or low education.

The result of the logistic regression using interaction variables was a model with an improved closeness of fit from the Hosmer and Lemeshow test (Chi-square = 0.909, *df* = 5, Sig. = 0.970). However, five of the six independent variables were not significant. The only significant variable (p=0.008) was the interaction variable combining education and income. The income and education interaction variable yielded an odds ratio of 1.719 (1.149-2.571, 95% C.I.). Table 5.4 describes the interaction variables and independent variables used in this stage of the analysis.

The logistic regression was run a second time using income-education interaction variable, the income variable, and the education variable as independent variables. The education variable proved significant, and upon combining the income variable regression coefficient with the income-education regression coefficient, the odds ratio for the education portion of the education-income education was 3.543, which was close to the risk estimate in the univariate and multivariate models.

TABLE 5.4

Interaction and Standard Variable Detail

Variable	Type	B	df	p	Odds Ratio	95% C.I.	
						Low	High
Urban-Education	Interaction	0.105	1	0.622	1.111	0.731	1.687
Urban-Income	Interaction	0.341	1	0.113	1.406	0.922	2.143
Income-Education	Interaction	0.541	1	0.008	1.719	1.149	2.571
Income	Standard	-0.397	1	0.265	0.672	0.334	1.352
Education	Standard	0.236	1	0.634	1.266	0.479	3.345
Urban	Standard	0.555	1	0.067	1.743	0.962	3.157
Clusters	Constant	-1.603	1	0.006	0.201	-	-

Relationship Between Prostate Cancer Incidence and Socioeconomic Status

The first stage of the statistical analyses took into account tracts with clusters as the independent variable. In the second stage, clusters were not considered in determining socioeconomic risk factors for prostate cancer incidence. Instead, the number of cases and the population in each age-race-ethnicity category for each individual tract was considered in a Poisson distribution analysis, as described in Chapter three.

The output of the Poisson analysis showed that there were nine parameters in the model. The nine parameters include the constant calculated from the cases and population, the four recoded quartiles of household income, and the four quartiles of education. The lowest education quartile and the lowest income quartile variables were identified as redundant and their parameter estimates were set to zero.

The goodness of fit statistics did not show that the Poisson loglinear model fit the data well. The likelihood ratio goodness of fit statistic resulted in a chi-square value of 21,592.4654 with nine degrees of freedom and a low significance value (p=0.000). The Pearson goodness of fit statistic resulted in a chi-square value of 19,493.6018 with nine degrees of freedom and a low significance value (p=0.000). Table 5.5 shows the results of the initial Poisson analysis.

In an effort to improve the goodness of fit, the variables were transformed as they were in the logistic regression. The socioeconomic variables were compressed from four quartiles into two groups each.

The output of the Poisson analysis with collapsed variables showed that there were five parameters in the model. The five parameters include the constant calculated from the cases and population, the two recoded values of household income, and the two recoded

TABLE 5.5

Initial Poisson Analysis Results

Variable	Std. Err.	Z-Value	Estimate	95% C.I.	
				Low	High
Cases (Constant)	0.0156	8.12	0.1268	0.10	0.16
Education High	0.0147	68.40	1.0084	0.98	1.04
Education Med-High	0.0152	32.89	0.4993	0.47	0.53
Education Med-Low	0.0166	1.99	0.0331	0.00042	0.07
Education Low	-	-	0.0000	-	-
Income High	0.0146	-12.57	-0.1831	-0.21	-0.15
Income Med-High	0.0153	-20.80	-0.3182	-0.35	-0.29
Income Med-Low	0.0160	-19.32	-0.3091	-0.34	-0.28
Income Low	-	-	0.0000	-	-

values of education. The lowest half education and the lowest half income variables were identified as redundant and their parameter estimates were set to zero.

The transformation yielded significantly better goodness of fit statistics, yet the goodness of fit was still not acceptable. The likelihood ratio goodness of fit statistic resulted in a chi-square value of 14,571.5489 with one degrees of freedom and a low significance value (p=0.000). The Pearson goodness of fit statistic resulted in a chi-square value of 13,669.1461 with one degrees of freedom and a low significance (p=0.000).

One reason for the poor fit of the model may be the collinearity of the variables education and income. To adjust for this problem, the variables were run independently in two separate Poisson analyses. The results of this Poisson test indicated similar results to the cluster regression results. All four variables in the Poisson analysis were significant. Table 5.6 shows the results of the Poisson analysis including the transformed variables.

TABLE 5.6

Transformed Variable Poisson Analysis Results

Variable	Standard Error	Z-Value	Odds Ratio Estimate	95% C.I.	
				Low	High
Education High	0.0083	250.92	2.0840	2.07	2.10
Education Low	0.0103	38.53	0.3967	0.38	0.42
Income High	0.0079	283.81	2.2526	2.24	2.27
Income Low	0.0101	12.78	0.1290	0.11	0.15

Conclusions

The purpose of this chapter of the research was to determine possible relationships between prostate cancer clusters, prostate cancer incidence, socioeconomic status, and rural place of residence.

Using existing prostate cancer incidence data and the clusters discovered in Chapter three, conclusions were drawn about the relationships between prostate cancer incidence, socioeconomic status, and rural place of residence. Logistic regression analysis and Poisson distribution analysis are standard epidemiologic approaches to analyzing these relationships and were used here to draw meaningful conclusions.

The logistic regression showed that living in a tract with high education, high income, or in an urban area does significantly predict a cluster of prostate cancer incidence. The range of risk estimates for the higher education and income groups as well as urban place of residence were significantly positive when not entered into a model with other variables (OR=3.631-5.936). In a logistic model containing income, education, and place of residence, odds ratios were all slightly lower than in the univariate models (OR= 2.153-3.306).

Interaction variables between the three independent variables produced a more desirable fit to the model, yet only the education-income interaction variable proved significant. High education and high income yielded an adjusted odds ratio of 3.543, which was higher than the odds ratio for high education alone (OR = 3.306) in the income-education-place of residence model. This result might indicate that education and income are more important together than separate in the model. The result also indicates that the socioeconomic factors together are more significant than urban place of residence

as part of an interaction variable in predicting tracts with clusters of prostate cancer incidence.

The Poisson analysis showed that high education and high income are both significant predictors of developing prostate cancer when adjusting for age, race, and ethnicity. High education had an odds ratio of 2.0840 (2.07-2.10, 95% C.I.) and high income had an odds ratio of 2.2526 (2.24-2.27, 95% C.I.).

The fact that the higher socioeconomic levels predicted prostate cancer clusters and incidence may be attributed to several reasons. The first reason might be that higher socioeconomic classes have more of a propensity to be screened for prostate cancer. As described in Chapter 2, many cases of prostate cancer go undiagnosed in patients if they do not receive screening, since there are no obvious symptoms in the early stages of the disease. Those with higher education may be more informed of the latent nature of the disease and begin screening at an earlier age and with more frequency. Those with higher income would likely have more health insurance and be more likely to be screened as well. Thus, higher income and education might not have anything to do with *developing* the disease. The scenario could be that men with higher education and income have a greater chance of reporting cases of prostate cancer incidence to the Texas Department of Health.

The second possible reason for the increased risk of developing prostate cancer for the higher socioeconomic classes might be environmental exposures or risk factors associated to lifestyle. Socioeconomic status may be an indicator or other environmental exposures that might be risk factors for prostate cancer incidence including diet, occupation, physical conditioning, and hazardous chemical exposures, among others.

In both types of tests, socioeconomic factors significantly predicted either prostate cancer clusters or incidence. These results suggest that clusters of prostate cancer, when adjusted for age-race-ethnicity, may be used as an independent variable in risk factor regression analysis of prostate cancer incidence. Conversely, conclusions about risk factors for prostate cancer clusters might be drawn from prostate cancer incidence analysis.

Although the results of both the cluster and incidence rate analyses were similar, one disparity was the difference in the effects of the various factors. In the cluster analysis, high education had a greater odds ratio (OR = 3.306) than high income (OR = 2.153). The results of the incidence analysis (clusters were not considered) showed that high income had a *slightly greater* odds ratio (OR = 2.2526) than education (OR = 2.0840). This result is important because it defines a disparity in the two methods of analyzing prostate cancer incidence. One important result of this thesis is the interchangeability of prostate cancer cluster and incidence rates as dependent variables when analyzing socioeconomic risk factors.

Several future research directions can be identified within the framework of this chapter. The focus of the first future research direction is on the number of socioeconomic variables used in the research. An expanded research might include several other socioeconomic indicators *in addition to* education and income. The use of only two socioeconomic variables significantly affected the closeness of fit of the models and made conclusions less concrete.

The focus of the second future research direction is the size of the area being examined. There is too much variation in socioeconomic factors across a census tract and

a smaller unit of geographical area, such as a census block group, would be desirable. This smaller geographic area would yield a more precise representation of socioeconomic status for each case in the research.

Another future research direction concerns obtaining more accurate population data with information such as migration and updated socioeconomic conditions. Socioeconomic status data from the 2000 census was not available for this study. Over a period of 10 years, it is possible for an area to change significantly in terms of socioeconomic conditions.

The final future research direction would be to incorporate potential environmental risk factors such as hazardous waste sites, air pollution, and other point sources of possible risk into the analysis. Adding an environmental component to the research would significantly alter the methodology, but is necessary to test for other possible confounding risk factors in addition to socioeconomic conditions and place of residence.

CHAPTER 6

CONCLUSIONS AND CONTRIBUTIONS

Conclusions

Cluster analysis has long been considered essential in the field of spatial epidemiology. However, many studies in the literature begin and end with cluster analysis. Cluster studies should be pre-epidemiology: analytic investigations that are done prior to more traditional, time-consuming, and costly epidemiologic designs (Albert, Gesler, and Levergood 2000). Most studies fail to integrate causation analysis with cluster analysis. This research is unique in the fact that it not only searches for clusters, but it continues to examine possible socioeconomic and geographical causes of clusters and incidence of prostate cancer. First, cluster analysis was used to detect clusters of prostate cancer incidence at the census tract level in Texas from 1990-1997. Further, logistic regression and Poisson regression were used to examine environmental risk factors for clusters and rates of prostate cancer incidence.

Research Questions

This thesis addressed four primary questions: (1) Are there any statistically significant spatial clusters of prostate cancer incidence in Texas at the census tract level?; (2) is there any statistically significant association between prostate cancer incidence

clusters and socioeconomic status at the census tract level in Texas?; (3) is there any significant association between prostate cancer incidence clusters and rural place of residence in Texas?; and (4) is socioeconomic status a significant risk factor for prostate cancer incidence in Texas at the census tract level?

The first research question in this thesis was related to spatial cluster analysis. Kulldorff's spatial scan statistic (Kulldorff 1997) was applied to population, geographic, and cancer incidence data from 1990-1997 to perform the cluster analysis. The null hypothesis stating that significant clusters of prostate cancer incidence do not exist was rejected. A statistically significant most likely cluster of prostate cancer incidence was found, located in north Bexar County, and comprising of 32 census tracts. The research also found 23 additional secondary statistically significant clusters in various sections of Texas.

The second question was related to socioeconomic factors that might influence clusters of prostate cancer incidence. The two socioeconomic factors examined include median household income and percentage of population over age 25 with some college education. The null hypothesis that socioeconomic factors do not predict clusters of prostate cancer incidence at the tract level in Texas was rejected. High levels of education and income proved to be statistically significant predictors of prostate cancer incidence clusters. Income and education together as an interaction variable also significantly predicted clusters of prostate cancer incidence.

The third research question was related to rural or urban place of residence as indicators of tracts containing clusters of prostate cancer. The null hypothesis that place of residence is not a predictor of clusters of prostate cancer incidence at the census tract

level in Texas was rejected here, as well. Urban place of residence was a statistically significant predictor of census tracts containing clusters of prostate cancer incidence to the approximate degree that income and education were predictors.

The fourth research question was related to socioeconomic factors that predict prostate cancer incidence. The null hypothesis that socioeconomic factors are not significant risk factors for prostate cancer incidence at the census tract level in Texas was rejected. Education and income were both significant risk factors for prostate cancer. There was quite a disparity between the respective odds ratios for income and education, though. Education had over six times the odds ratio of income.

The results of this thesis contribute to studies in the literature of a similar nature. A review of cancer research by Dale et al. (1996) showed that low education and low income were risk factors for prostate cancer mortality in most studies. This research, however, found that *high* income and *high* education were significant risk factors for developing prostate cancer. The difference could be attributed to the fact that this research focused on incidence rather than mortality. As mentioned in Chapter 5, incidence levels might be elevated in the higher socioeconomic classes due to the lack of testing of men living in lower socioeconomic conditions.

Further, this research found a significant correlation between urban place of residence and prostate cancer incidence clusters and rates. The limited amount of literature available was focused on mortality and resulted in higher rates among those living in rural locations, such as farmers. The contrary evidence presented in this research may provide additional insight into urban place of residence as a risk factor for prostate cancer incidence as opposed to rural place of residence.

Contributions

This research contributes to several different theories and bodies of knowledge including the field of spatial epidemiology and human ecology theory.

This research has contributed to the knowledge on spatial analysis of prostate cancer incidence. The spatial distribution of prostate cancer incidence was studied in depth. This prostate cancer incidence cluster research is the first of its kind at the census tract level in Texas. The most likely cluster in North San Antonio and the other secondary clusters scattered across the state are valuable starting points for research investigating possible environmental risk factors.

This thesis also contributed to the social aspects of environmental spatial data analysis. Social status has long been an indicator of one's health and longevity. This thesis contributes to the social aspects of environmental analysis by relating one's place of residence and social status to clusters and rates of prostate cancer incidence. These findings are useful because they disclose how people's residence locations and socioeconomic status affect their health conditions (Baquet et al. 1991). This thesis reveals helpful hints for future public health planning and management officials about the location and nature of prostate cancer clusters and incidence.

Human ecology theory focuses on the interaction between individuals and their environment. In this thesis, individual cases of prostate cancer incidence were attributed to environmental indicators based on socioeconomic status. Socioeconomic status was shown to have a statistically significant impact on both clusters of prostate cancer incidence and prostate cancer incidence rates. Thus, this research supports human ecology theory as evidence of the theory's validity.

Final Thoughts

What causes prostate cancer? Where does the cure lie? These are questions that have eluded researchers for years, and will, undoubtedly, persist into the near and perhaps distant future. In the final analysis, this research does not provide any magical framework for answering these critical questions. It does, however, contribute a few ounces of understanding to the well of cancer knowledge. How deep that well is, no one knows. We must continue to pour into that well, little by little, until the water level raises enough for us to observe the answers in our own reflections.

REFERENCE LIST

Albert, D. P., W. M. Gesler, and B. Levergood. 2000. *Spatial analysis, GIS and remote sensing applications in the health sciences.* Chelsea, MI: Ann Arbor Press.

American Cancer Society. 2001. Available from http://www.cancer.org/; Internet; accessed 06 November 2001.

Bailey, Trevor C. 1994. A review of statistical spatial analysis in geographical information systems. In *Spatial analysis and GIS,* eds. A. Stewart Fotheringham and Peter A. Rogerson, 13-44. London: Taylor & Francis.

Baquet, C. R., J. W. Horm, T. Gibbs, and P. Greenwald. 1991. Socioeconomic factors and cancer incidence among blacks and whites. *Journal of the National Cancer Institute* 83, no. 8: 551-57.

Bertalanffy, Ludwig von. 1969. General systems theory: Foundations, developments and applications. New York: G. Braziller.

Brender, Jean. 2001a. Lecture and class material. Health Research 5351. *Principles of Epidemiology.* 29 August. Southwest Texas State University.

_____. 2001b. Lecture and class material. Health Research 5351. *Principles of Epidemiology.* 05 September. Southwest Texas State University.

_____. 2001c. Lecture and class material. Health Research 5351. *Principles of Epidemiology.* 19 September. Southwest Texas State University.

Bullard, Robert D. 1990. *Dumping in Dixie: Race, class, and environmental quality.* Boulder, CO: Westview Press.

Bunge, W. 1962. *Theoretical geography*. Lund, Sweden: Series C Publishing.

Carstairs, V. 2000. "Socio-economic factors at areal level and their ralationship with health." In *Spatial Epidemiology: Methods and Applications*, eds. P. Elliott, J. C. Wakefield, N. G. Best, and D. J. Briggs. Oxford, U.K.: Oxford University Press.

Carter, Bob S., H. B. Carter, and J. T. Isaacs. 1990. Epidemiologic evidence regarding predisposing factors to prostate cancer. *The Prostate* 16, no. 2: 187-97.

Cella, D. F., E. J. Orav, A. B. Kornblith, J. C. Holland, P. M. Silberfarb, K. W. Lee, R. L. Comis, M. Perry, R. Cooper, L. H. Maurer, D. F. Hoth, M. Perloff, C. D. Bloomfield, O. R. McIntyre, L. Leone, G. Lesnick, N. Nissen, A. Glicksman, E. Henderson, M. Barcos, R. Crichlow, C. S. Faulkner II, W. Eaton, W. North, P. S. Schein, F. Chu, G. King, and A. P. Chahinian.. 1991. Socioeconomic status and cancer survival. *Journal of Clinical Oncology* 9, no. 8: 1500-509.

Chakraborty, Jayajit, and Marc P. Armstrong. 2001. Assessing the impact of airborne toxic releases on populations with special needs. *Professional Geogrpaher* 53, no. 1: 119-31.

Coleman, J.S. 1964. *Introduction to mathematical society.* New York: Free Press.

Crocetti, Emanuele, Stefano Ciatto, and Marco Zappa. 2001. Prostate cancer: Different incidence but not mortality trends in two areas of Tuscany, Italy. *Journal of the National Cancer Institute* 93, no. 11: 876-77.

Cytel Software Corporation. 1999. *Egret users manual for windows: software for the analysis of biomedical and epidemiological studies*. Cambridge, MA: Cytel Software Corporation.

Dale, W., S. Vijayakumar, E. F. Lawlor, and K. Merrell. 1996. Prostate cancer, race, and socioeconomic status: Inadequate adjustment for social factors in assessing racial differences. *The Prostate* 29, no. 3: 271-81.

Dayal, H. H. and C. Chui. 1982. Factors associated with racial differences in survival for prostatic carcinoma. *Journal of Chronic Diseases* 35, no. 5: 553-60.

Dayal, H. H., L. Polissar, and S. Dahlberg. 1985. Race, socioeconomic status and other prognostic factors for survival from prostate cancer. *Journal of the National Cancer Institute* 74, no. 5: 1001-006.

Deonandan, Raywat, K. Campbell, T. Ostbye, I. Tummon, and J. Robertson. 2000. A comparison of methods for measuring socioeconomic status by occupation or postal area. *Chronic Disease in Canada* 21, no. 3: 123-28.

Dijkman, G. A., and F. M. J. Debruyne. 1996. Epidemiology of prostate cancer. *European Urology* 30, no. 2: 281-95.

Ernster, V. L., S. Selvin, S. T. Sacks, D. F. Austin, S. M. Brown, and W. Winderstein. 1978. Prostatic cancer: Mortality and incidence rates by race and social class. *American Journal of Epidemiology* 107, no. 4: 311-20.

Elliott, P., J. Cuzick, D. English, and R. Stern. 1996. *Geographical and environmental epidemiology: Methods for small-area studies.* New York: Oxford University Press.

Elliott, P., J. C. Wakefield, N. G. Best, and D. J. Briggs. 2000. *Spatial epidemiology: Methods and applications.* Oxford: Oxford University Press.

Environmental Protection Agency. 2002. Available from http://www.epa.gov; Internet; accessed 15 February 2002.

Fischer, M. M. 1999. Spatial analysis: Retrospect and prospect. In *Geographical Information Systems,* 2d ed., eds. P. A. Longley, M. F. Goodchild, D. J. Maguire, and D. W. Rhind, 283-92. John Wiley and Sons.

Gatrell, Anthony. 1983. *Distance and space: A geographical perspective.* Oxford, U.K.: Clarendon Press.

Haining, Robert P. 1998. Spatial statistics and the analysis of health data. In *GIS and Health,* eds. A. C. Gatrell and M. Löytönen, 29-47. London: Taylor and Francis.

Hanchette, Carol L., and Gary G. Schwartz. 1992. Geographic patterns of prostate cancer mortality. *Cancer* 70, no. 12: 2861-869.

Hardy, R., and M. K. Hargreaves. 1991. Cancer prognosis in black Americans: A mini-review. *Journal of the National Medical Association* 83, no. 7: 574-79.

Higginson, John. 1983. The face of cancer worldwide. *Hospital Practice* 18, no. 11: 145-57.

Hjalmars, U., M. Kulldorff, G. Gustafsson, and N. Nagarwalla. 1996. Childhood leukemia in Sweden: Using GIS and a spatial scan statistic for cluster detection. *Statistics in Medicine* 15, no. 5: 707-15.

Hoffman, R. M., F. D. Gilliland, J. W. Eley, L. C. Harlan, R. A. Stephenson, J. L. Stanford, P. C. Albertson, A. S. Hamilton, W. C. Hunt, and A. L. Potosky. 2001. Racial and ethnic differences in advanced-stage prostate cancer: The prostate cancer outcomes study. *Journal of the National Cancer Institute* 93, no. 5: 388-95.

Holifield, Ryan. 2001. Defining environmental justice and environmental racism. *Urban Geography* 22, no. 1: 78-90.

Holt-Jensen, Arild. 1988. *Geography: History and concepts.* 2nd ed. Totowa, N.J.: Barnes & Noble Books.

Hsing, A., and S. Devesa. 2001. Trends and patterns of prostate cancer: What do they suggest? *Epidemiologic Reviews* 23, no. 1: 3-13.

Jacquez, Geoff, and Leah Estberg. 2001. *ClusterSeer: Software for identifying disease clusters, a user guide.* Ann Arbor, MI: TerraSeer.

Johns Hopkins Oncology Center. 2001. Available from http://www.hopkinscancercenter.org/ types/prostate.cfm; Internet; accessed 07 November 2001.

Jones, J. 2001. News: African-Americans and prostate cancer: Why the discrepancies? *Journal of the National Cancer Institute* 93, no. 5: 342-44.

Kelada, S. N., S. L. R. Kardia, A. H. Walker, A. J. Wein, S. B. Malkowicz, and T. R. Rebbeck. 2000. The glutathione s-transferase-μ and –θ genotypes in the etiology of prostate cancer: Genotype-environment interactions in smoking. *Cancer Epidemiology, Biomarkers and Prevention* 9, no. 12: 1329-334.

Knox, Paul. 1982. Urban social geography. New York: John Wiley & Sons, Inc.

Kulldorff, Martin. 1997. A spatial scan statistic. *Communication in Statistics: Theory and Methods* 26, no. 6: 1481-496.

_____. 1998. Statistical methods for spatial epidemiology: Tests for randomness. Chapter 4 of *GIS and Health*, ed. Ian Masser and Francois Salge, 49-62. London: Taylor & Francis Ltd.

Kulldorff, M., and N. Nagarwalla. 1995. Spatial disease clusters: Detection and inference. *Statistics in Medicine* 14, no. 4: 799-810.

Kulldorff, Martin, Katherine Rand, Greg Gherman, Gray Williams, and David DeFrancesco. 1998. *SaTScan v2.1: Software for the spatial and space-time scan statistics.* Bethesda, MD: National Cancer Institute.

Last, J. M. 2001. *A dictionary of epidemiology.* 4th ed. New York: Oxford University Press.

Martin, Geoffrey J., and Preston E. James. 1993. *All possible worlds: A history of geographical ideas.* 3rd ed. New York: John Wiley & Sons, Inc.

McNeil, Don. 1996. *Epidemiological research methods.* New York: John Wiley & Sons.

McWhorter, W. P., A. G. Schatzkin, J. W. Horm, and C. C. Brown. 1989. Contribution of socioeconomic status to black/white differences in cancer incidence. *Cancer* 63, no. 5: 982-87.

Meade, M. S., and R. J. Earickson. 2000. *Medical geography.* New York: The Guilford Press.

Mebane, C., T. Gibbs, and J. Horm. 1990. Current status of prostate cancer in North American black males. *Journal of the National Medical Association* 82, no. 11: 782-88.

Mertler, Craig A. and Rachel A. Vannatta. 2002. *Advanced and multivariate statistical methods: Practical application and interpretation.* Los Angeles, CA: Pyrczak Publishing.

National Center for Health Statistics. *Health, United States, 1998 with socioeconomic status and health chartbook.* Hyattsville, MD: National Center for Health Statistics.

_____. *Health, United States, 1999 with health and aging chartbook.* Hyattsville, MD: National Center for Health Statistics.

National Institutes of Health. 2001. Available from http://grants.nih.gov/grants/guide/rfa-files/RFA-AG-02-003.html; Internet; accessed 10 October 2001.

Park, Robert E., Ernest W. Burgess, and Roderick D. McKenzie. 1825. The City. Chicago, IL: University of Chicago Press.

Poisson, S. –D. 1837. *Recherches sur la probabilité des jugements en matière criminelle et en matière civile.* Paris: Bachelier.

Polednak, A. P. 1990. Cancer mortality in a higher-income black population in New York state: Comparison with rates in the United States as a whole. *Cancer* 66, no. 7: 982-87.

_____. 2001. Association of African-American ethnic background with survival in men with metastatic prostate cancer. *Journal of the National Cancer Institute* 93, no. 15: 1174-75.

Potosky, A. L., E. J. Feuer, and D. L. Levin. 2001. Impact of screening on incidence and mortality of prostate cancer in the United States. *Epidemiologic Reviews* 23, no. 1: 181-86.

Principia Cybernetica. 2002. Available from http://pespmc1.vub.ac.be; Internet; accessed 28 February 2002.

Risser, David. 2001. Texas Department of Health. Personal communication via email by author. Southwest Texas State University. 5 December.

_____. 2001b. Texas Department of Health. Personal communication via email by author. Southwest Texas State University. 18 December.

Ross, R. K., J. W. McCurtis, B. E. Henderson, H. R. Menck, T. M. Mack, and S. P. Martin. 1979. Descriptive epidemiology of testicular and prostatic cancer in Los Angeles. *British Journal of Cancer* 39, no. 2: 284-92.

Ross, R. K., and D. Schottenfeld. 1996. Prostate Cancer. In *Cancer epidemiology and prevention*, eds. David Schottenfeld, and Joseph E. Fraumeni, Jr. New York: Oxford University Press.

Shibata, A. and A. S. Whittemore. 1997. Genetic predisposition to prostate cancer: Possible explanations for ethnic differences in risk. *The Prostate* 32, no. 1: 65-72.

Snow, John. 1854. *On the mode of communication of cholera*, 2nd ed. London: John Churchill, New Burlington Street.

Selvin, Steve. 1996. *Statistical analysis of epidemiologic data.* New York: Oxford University Press.

SPSS. 1999. *SPSS advanced models 10.0.* Chicago: SPSS, Incorporated.

SPSS. 2000. *SPSS for windows version 10.0.7.* Chicago: SPSS, Incorporated.

The Daily University Star. 1998. Cancer remains everyday reality for Supple; Available from http://www.star.swt.edu/98/10/01/news.html#news6; Internet; accessed 1 December 2001.

Texas Department of Health. 1999. Prostate cancer advisory committee: 1998-1999 biennial report. Austin, TX: Texas Department of Health.

_____. 2001. Available from http://www.tdh.state.tx.us/osp/prostate.htm; Internet; accessed 7 November 2001.

Texas Department of Health Texas Cancer Registry. 2001. Available from http://www.tdh.state.tx.us/tcr/default.html; Internet; accessed 6 November 2001.

Texas State Data Center. 2001. Texas A&M University. Available from http://txsdc.tamu.edu/; Internet; accessed 1 November 2001.

Timmreck, T. C. 1994. *An introduction to epidemiology.* Boston: Jones and Bartlett Publishers.

United States Census Bureau. 2001. Available from http://www.census.gov/; Internet; accessed 5 December 2001.

University of North Texas. 2001. Available from http://www.unt.edu/cpe/module2/thrybase.htm; Internet; accessed 28 February 2001.

Yu, H., R. E. Harris, and E. L. Wynder. 1988. Case control study of prostate cancer and socioeconomic factors. *Prostate* 13, no. 3: 317-25.

Zhan, F. B. 2001. Childhood cancer clusters in New Mexico, 1973-1997. *The Southwestern Geographer,* (in press).

_____. 2002. Are mortality cases of liver cancer, lung cancer, kidney cancer, and leukemia clustered in San Antonio? *Texas Medicine*, (in press).

Zhou, Xinnong. 2000. Geographic concentrations of lung cancer mortality in Texas and their relationship to environmental and socioeconomic conditions. Ph.D. diss., Soutwest Texas State University.

VITA

Jeffrey Gaines Wilson was born in Austin, Texas, on July 31, 1973, the son of Art and Kathryn Wilson. After graduating in 1992 from Boerne High School, Boerne, Texas, he entered St. Gregory's College in Shawnee, Oklahoma. His sophomore year, he transferred to Texas A&M University in College Station and graduated with the degree of Bachelor of Science in May of 1996. Jeffrey spent the next four years working for various public and private corporations, including International Business Machines Corporation and Ernst & Young LLP. Jeffrey returned to Texas to study spatial epidemiology under Dr. Ben Zhan at Southwest Texas State University. He received the degree of Master of Science in May of 2002. Jeffrey has been accepted into the Ph.D. program at the University of Canterbury on the South Island of New Zealand, where he will continue his research in spatial epidemiology.

Permanent address: Post Office Box 1754
Boerne, Texas 78006-6754
Unites States of America

This thesis was typed by Jeffrey Gaines Wilson.