A SPATIAL ANALYSIS OF INVASIVE BREAST CANCER CLUSTERS IN ASSOCIATION WITH ENVIRONMENTAL RISK FACTORS: ILLINOIS 1996 TO 2000

by

William C. Weston

B.A., Southern Illinois University Carbondale, 2005

A Thesis
Submitted in Partial Fulfillment of the Requirements for the
Master of Science Degree

Department of Geography and Environmental Resources
In the Graduate School
Southern Illinois University Carbondale
August 2012

THESIS APPROVAL

A SPATIAL ANALYSIS OF INVASIVE BREAST CANCER CLUSTERS IN ASSOCIATION WITH ENVIRONMENTAL RISK FACTORS: ILLINOIS 1996 TO 2000

by

William C. Weston

A Thesis Submitted in Partial

Fulfillment of the Requirements

for the Degree of

Master of Science

in the field of Geography and Environmental Resources

Approved by:

Dr. Tonny J. Oyana, Chair

Dr. Dhitinut Ratnapradipa

Mr. Samuel Adu-Prah

Graduate School
Southern Illinois University Carbondale
May 21, 2012

AN ABSTRACT OF THE THESIS OF

WILLIAM C. WESTON, for the Master of Science Degree in GEOGRAPHY AND ENVIRONMENTAL RESOURCES, presented on May 21st, 2012, at Southern Illinois University Carbondale.

TITLE: A SPATIAL ANALYSIS OF INVASIVE BREAST CANCER CLUSTERS IN ASSOCIATION WITH ENVIRONMENTAL RISK FACTORS: ILLINOIS 1996 TO 2000

MAJOR PROFESSOR: Dr. Tonny J. Oyana

This retrospective study assesses invasive breast cancer counts reported at the Illinois ZIP code scale during the study period of 1996 to 2000. The research objective is to evaluate the spatial and statistical associations between breast cancer risk and sources of potential environmental contamination. A thorough literature review illustrates a profound list of cancer risk factors within the study space. Public health principles are utilized to prepare breast cancer incidence for analysis, accompanied with the development of a case/control ecological model. Exploratory analyses suggest that breast cancer intensity is predominantly a rural problem. A generalized linear mixed model is employed, illustrating statistical associations between environmental risk factors and breast cancer risk. Coal Mines, Oil/Gas Wells, and Large Quantity Hazardous Waste Generators, display high statistical significance ($p<0.001$) in association with increased breast cancer risk. Unique socioeconomic attributes distinguish urban risk from rural risk, as can be seen in a discriminant function analysis. The modeling techniques utilized in this research display classic spatial epidemiological approaches that account for particular types of confounding effects, while also defining zones of disease risk through cluster detection. Results from this analysis are useful for future studies intended to account for epidemiological, clinical, chemical and biological disease-related information.

ACKNOWLEDGEMENTS

I would like to thank Dr. Tonny J. Oyana for being a person of principle, consistency and scientific vision. His subject matter expertise in spatial epidemiology has provided mountains of wisdom and effective guidance.

My gratitude is also extended to Mr. Samuel Adu-Prah and Dr. Dhitinut Ratnapradipa, whose commitment to the academic experience is profoundly humbling. I am very grateful for their focused words and scientific guidance.

I am further grateful for the research that previous students have come before me to perform. Those works have provided scientific guidance and perspective during the construction of this research. It is my hope that many more students will follow this well-lit path and employ GIS and spatial analysis within epidemiologic research.

TABLE OF CONTENTS

CHAPTER PAGE

LIST OF TABLES

TABLE PAGE

LIST OF FIGURES

FIGURE PAGE

FIGURE PAGE

CHAPTER 1

INTRODUCTION

1.1 Background

In the United States, 12.15 percent of the female population (or one out of eight females) is expected to battle breast cancer at some point during life (Jemal 2010; SEER 2010a). An estimated 209,060 new cases of invasive breast cancer were estimated for the year 2010, with a minor constituent frequency of 1,970 cases occurring in males (ACS 2010). Breast cancer is the first and second leading cause of cancer-related death in women ages 40-59 and 60-79, respectively (Jemal 2010). Demographically, US women who report as White-Hispanic, Asian/Pacific Islander, Black, and White non-Hispanic accounted for the four highest breast cancer risk ethnicities, with rates of 92.7, 97.2, 121.7 and 148.3 cases per 100,000 women, respectively (SEER 2010b). Wealthier socioeconomic status has also been shown to express higher breast cancer morbidity (Brody and Rudel 2003).

Nationally, from 1996 to 2000, the US age and population-adjusted breast cancer rate for all women was 137.1 cases per 100,000 women (SEER 2010b), slightly higher than the Illinois breast cancer rate of 133.3 cases per 100,000 women (Dolecek *et al.* 2003). From 2002 to 2006, Illinois' female breast cancer incidence rate (123.1 cases per 100,000 women) was slightly above the US female breast cancer rate of 121.8 cases per 100,000 women. At the same time, Iowa was reported as the only state in the central US to exceed Illinois' breast cancer incidence (ACS 2010). Approximately sixty-eight (68) percent of Illinois females report as White non-Hispanic (US Census Bureau 2010a), and 44 percent of Illinois females report as a

member of the 40 and older age group (Census Scope 2010), placing a large percentage of Illinois women at higher socioeconomic and intrinsic risk for breast cancer.

A common problem inherent to epidemiologic studies of environmental cancer risk is that environmental exposure measures do not necessarily reveal etiology (Brody and Rudel 2003). This is plausible given the myriad of factors that serve etiologically in breast cancer incidence. More discontented researchers position themselves against epidemiologic studies of environmental risk, claiming that epidemiologic science fails to serve its purpose when it associates cancer risk with environmental exposure to man-made chemicals such as endocrine disrupting chemicals (Ames and Gold 2000).

During the past decade, a number of cancer epidemiologists have brought to light many of the suspected environmental risk factors (i.e. volatile organic compounds, inorganic compounds, radionuclide isotopes, and heavy metals) that operate as carcinogens or endocrine disrupting factors within different geographic populations (Birnbaum and Fenton 2003; Brody and Rudel 2003; Lichtenstein 2000; Rudel *et al.* 2007). 216 potential mammary carcinogens have been identified via recent studies on animals, in addition to 250 estrogen mimics (resembling human body hormones) in animals (Brody *et al.* 2007).

A spatial epidemiological focus on environmental risk factors does take into consideration -albeit limited at times- risk factors such as lifestyle behaviors, hereditary dynamics, and other individually isolated space-time variables. To better understand the nature of a cancer epidemic, it is necessary to understand how individuals interact with their spatial environments as individual organisms. In the case of environmental risk, this study adheres to the perspective that people are frequently subjected to exposure from environmental phenomena such as groundwater contamination, soil and air pollution, and local industrial activities. These

are exposure pathways for which most communities cannot control without major restorative assistance or government regulation.

In essence, everything shares a spatial relationship with everything, and more approximate objects tend to relate more with each other (Tobler 1979). I argue that individuals who are in closer proximity to cancer causing agents will in turn experience higher causal effects from those carcinogenic substances. This type of perspective perceives disease outcomes as phenomena that are dependent upon the geographic scales at which they are observed –with variations in scale associating with variations in observed disease outcomes (Moore and Carpenter 1999). Moore and Carpenter (1999) also note that spatial analysis and GIS have many current uses in epidemiology, including the investigation of environmental risk, detection of clusters, and assessments of distance dependent phenomena.

The total treatment costs of US cancer for the year 2006, estimated at $104.1 billion (NCI 2010a), juxtapose ironically with the fiscal year 2010 US Department of Defense budget authorization of $104.9 billion; this budget being used to support all fiscal military operations in Afghanistan (Belasco 2010). The predominantly allopathic medical industry of the US is in the midst of the largest [exponential] medical cost increase in history, displaying absolutely no tendency toward financial equilibrium (Cutler 2007; KFF 2010; NCI 2010a; WHO 2010).

Under current circumstances, approximately 37.76 percent of women and 44.29 percent of men in the US will acquire cancer at some point during life (SEER 2010a). This helps to explain the $263.8 billion in total direct and indirect cancer costs that the US was estimated to absorb in the year 2010 (ACS 2010). Perhaps, the US would be on a more economically sustainable and medically viable path if more resources were directed toward the prevention,

identification, and eradication of cancer risks prevalent in the US, especially in the midst of a wide-spread national recession (Inman 2010; Pincus-Nielsen *et al.* 2010).

On February 26th, 2009, it was considered a major improvement to Environmental Protection Agency (EPA) policy when the US President, Barack Obama, proposed a $3 billion increase in the Environmental Protection Agency's FY 2010 budget, moving the budget from $7.5 billion to $10.5 billion and including provisions for new environmental protections (EPA 2009). Despite the White House's new environmental health approach (PCP 2010; Cone 2010), all environmental agendas are minuscule in financial comparison to elevations in US healthcare expenditures that occurred from 2000 to 2006. During this time period, US per capita health expenditures increased 47 percent from $4,570 per person to $6,719 per person (WHO 2010). When factored into US population estimates (US Census Bureau 2010b), this 47 percent increase was equivalent to $716 billion in new health costs, bringing fiscal healthcare expenditures in the US to over $2 trillion by the end of 2006. Also in 2006, breast cancer treatment -the most expensive histologic site for cancer treatment- cost the US $13.88 billion (NCI 2010a), exceeding the US's modern EPA budget. Most of these phenomena suggest a systematic bias toward 'health symptom treatment' instead of 'health symptom prevention'.

Financial arguments against the cost of eliminating environmental risks, i.e. Ames and Gold (2000), fail to realize the financial benefits of a healthier population. As a result of lower lifetime pollutant exposure, fewer per capita dollars would be spent treating medical problems, and more per capita work hours would be spent productively in the US labor force. We must also consider the value of human life and love, both of which for many people exceed all things measured in dollars.

It is possible to preserve financial and emotional resources by identifying and reducing health endangering risks. Much work has been performed in the fields of spatial epidemiology and medical geography, and these efforts have helped with the better evaluation of environmental health.

1.2 Problem Statement

> "The overwhelming majority of chemicals identified as animal mammary carcinogens or endocrine-disrupting compounds have never been included in an epidemiologic study of breast cancer (Brody *et al.* 2007)."

The goal of this research is to determine the spatial and statistical associations that are shared between the locations of environmental risk factors and the locations of breast cancer risk. An emerging bedrock of evidence has revealed that many locations in Illinois are exposed to a wide field of environmental risk factors. The issue might be explainable by private and public drinking water supplies in Illinois, which are susceptible to both anthropogenic and naturally occurring contamination. Private drinking water supplies, which predominantly occur in rural areas, are not subject to regulations outlined in the US Safe Drinking Water Act (EPA 2010a). For this reason, private drinking water supplies are believed to be more vulnerable to carcinogenic and endocrine disrupting contamination. Public supplies pose their own degree of risk, given the potential for naturally occurring inorganic and radionuclide particulates to bypass municipal filtration.

I argue that through a more complete spatial and statistical understanding of environmental risk factor exposure, we will gain a better understanding of etiological associations with breast cancer risk in Illinois.

1.3 Research Questions

1) Which Illinois ZIP codes displayed breast cancer clustering during the study period?

2) Can environmental contaminants with a known or suspected etiologic role in breast cancer be found within Illinois?

3) Are breast cancer clusters associated with the locations of suspected environmental contaminants?

1.4 Significance of the Study

This research is an attempt to re-examine and build upon previous cancer research (Wang 2004) that illustrated a variety of cancer gradients (including breast cancer) in Illinois by ZIP code for the years of 1996 through 2000. Wang (2004)'s research did not associate environmental risk factors that might have influenced breast cancer outcomes. My intent is to re-evaluate data from this 1996-2000 timeframe and apply statistical analyses that measure spatial associations between incidence locations and environmental risk factor locations. This analysis will fill a gap in research by providing a better understanding of Illinois as a landscape that presents numerous breast cancer risks. It is also the intent of this research to identify –or at least empirically challenge- risk factors that could require remedial actions or stricter control measures.

I hypothesize that this study will reveal breast cancer risk to be higher in areas where exposure to environmental risk factors is prevalent. This hypothesis is presented as a

consequence of the older-age female demographic in Illinois and the wide array of environmental risk exposures that these women could have encountered throughout the latter half of the 20th century. This study has the potential to provide empirical insights to researchers and policy-makers who can influence environmental health protection, epidemiological intervention, and public health administration.

CHAPTER 2

LITERATURE REVIEW

2.1 Breast Cancer

Global estimates showed that in 2002, 1,152,000 women were diagnosed positively for breast cancer (equaling 37.4 percent of all global female cancer diagnoses), and 4,408,000 women were living with breast cancer; making breast cancer the leading histologic site for human cancer incidence and human cancer prevalence (Parkin *et al.* 2005). Without considering melanoma of the skin, breast cancer is the most common type of female cancer in the US, accounting for at least 1 in 4 cancer diagnoses in US females (ACS 2009; Parkin *et al.* 2005).

The incidence rate for breast cancer in the US has risen steadily at a rate of 1 percent per year since 1940 (Miller *et al.* 1994); however, much of that incidence data was obtained prior to standardized diagnosis procedures in the early 1970s. A dramatic increase in breast cancer occurred from 1980 to 2000, with a total increase of 33.3 percent. Rates increased from 102.22 to 136.28 cases per 100,000 women during this timeframe, peaking at 140.3 cases per 100,000 women in 1998 (SEER 2010). From 2002 to 2006, 95 percent of new breast cancer diagnoses and 97 percent of breast cancer deaths occurred in US women greater than 40 years of age (ACS 2009). Nearly 44 percent of the 6,338,957 women in Illinois comprise the age bracket of 40-85+ (Census Scope 2010), meaning that 2,769,464 women in Illinois are at highest risk for breast cancer morbidity and subsequent mortality. From 2003 to 2007, the incidence of 'in situ' and invasive breast cancer in Illinois increased 3.4 percent and 1.3 percent, respectively (NCI

2010c), revealing the need for increased medical screening, treatment, and prevention in Illinois. Although white women in the US have a higher incidence of breast cancer, black women in the US have a higher incidence among women before age 45 and are most likely to die from breast cancer at every age (Jemal 2010). From 1996 to 2000, the population-adjusted breast cancer rate in Illinois was 133.3 cases per 100,000 women, with white and black women showing rates of 134.3 and 121.9 cases per 100,000 women, respectively (Dolecek *et al.* 2003).

Pertinent to the period of 1996-2000, Wang (2004) performed a county level and ZIP code level analysis of cancer clusters. The ZIP code analysis revealed that higher gradients of breast cancer incidence occurred within the three primary metropolitan statistical areas (MSAs) of Illinois. This outcome reinforces the ability of ZIP code level data to reveal information that cannot be seen at lower resolutions.

2.2 Exposure to Environmental Cancer Risk

During a pesticide exposure assessment, the US EPA Federal Register 2005 (EPA 2007) illustrated four primary exposure pathways to pesticides. These pathways were identified as food, drinking water, residential use, and occupational contact. This master's research adopts a similar perspective where exposure to chemical contaminants can occur through an assortment of environmental pathways. The etiological sources of contaminant exposure could include water supply wells, agricultural chemical application areas, petroleum wells, waste-water release sites, power generation plants, hazardous waste generators, and numerous other sources. For example, a nuclear power plant could be responsible for releasing radioactive contaminants

into a local watershed, providing the opportunity for contaminants to enter local water supplies, which could then be consumed or utilized by women via drinking, bathing and cooking.

The Safe Drinking Water Act was passed by the US Congress in 1974 (and later amended in 1986 and 1996), in a nationwide effort to guard public health by allowing the US Environmental Protection Agency to govern the chemical constituents of municipal drinking water sources such as rivers, lakes, reservoirs, springs, and ground water wells, excepting private wells that serve less than 25 individuals (EPA 2010a). More than four million Illinois residents (approximately 35 percent of the state's population) utilize groundwater (private wells) as their primary water source; and, in rural areas, approximately 90 percent of Illinois residents utilize private wells to obtain their primary water supply (ISWS 2009a). The EPA does not have the authority to govern private wells, meaning that 15 percent of US wells (all private) are not subject to EPA regulation and are often unchecked by water experts for water contamination. However, these private wells could still be subject to state and local government regulation (EPA 2010a).

Since health officials are concerned with chronic health effects such as cancer, the US EPA's Primary Standards or Maximum Contaminant Levels (MCLs) for carcinogenic substances tend to be as close as possible to zero, without actually reaching zero (Stewart *et al.* 2001). Of utmost importance, Stewart *et al.* (2001) also stress that EPA Primary Standards do not guarantee that water with a contaminant level below the standard is risk free, just as they infer that water in excess of prescribed MCLs is not necessarily dangerous. MCLs are mostly the results of scientific estimates based on available information. It can be argued that environmental studies of cancer are direly needed, because the MCL enforcing authorities need concise information off of which to base their water contaminant policies.

2.3 Environmental Risk Factors

Predominant medical perspectives about breast cancer etiology support intrinsic factors such as age, genetics, family history, personal history, race, hormones, exercise, obesity, alcohol use and ionized radiation, as primary breast cancer risk factors (E-Medicine 2010; NCI 2010c; E Health MD 2010). However, growing numbers of environmentally aware communities (i.e. Breast Cancer Fund; *Environmental Health Perspectives* Journal; and President's Cancer Panel) have begun to consider the etiologic influence of environmental risk factors and how such risks play a role in cancer outcomes. Similarly, Lichtenstein *et al*. (2000)'s analysis of Swedish, Danish and Finish twin registries suggests that at least 60 percent of cancer incidence can be attributed to discrete, non-heritable, environmental factors. Such research suggests that intrinsic human factors are viable -but minor- etiologic associates of breast cancer incidence, prompting an influx in environmental risk research.

Carcinogenic substances are frequently referred to as 'mutagens', because they are capable of interrupting the chemical constituency and shape of deoxyribonucleic acid (DNA) molecules, promoting patterns of DNA replication that are preparatory for carcinogenesis (Campbell *et al.* 2008; Greenblatt *et al.* 1994). This is why carcinogenic substances are typically DNA mutagens, just as DNA mutagens are typically carcinogenic substances (Campbell *et al.* 2008). Volatile organic compounds (e.g. halogenated and polycyclic aromatic hydrocarbons from petroleum components) have been recognized for their carcinogenic and endocrine disrupting roles during mammary gland neoplasia (Brody and Rudel 2003; Rudel *et al.* 2007). Substances such as volatile and non-volatile organic compounds, inorganic compounds, radioactive isotopes, heavy metals, water disinfectants, hormones, and other

suspected chemicals, can function as carcinogens and/or endocrine disrupting agents when vertebrate animals (i.e. humans, rats, rabbits) encounter them (Brody and Rudel 2003; Brody *et al.* 2007; Birnbaum and Fenton 2003; Crisp *et al.* 1998; Ejaz 2004; EPA 2010b; Gray 2010; Rudel *et al.* 2007). Many of these substances act directly as DNA mutagens in animal tissues, but some operate indirectly as endocrine disruptors within the endocrine system, promoting alternate pathways and conditions for carcinogenesis (Birnbaum and Fenton 2003; Crisp *et al.* 1998). Brody and Rudel (2003) urge that endocrine disrupting compounds, through their ability to mimic the molecular shape of estrogen, can initiate gene expressions just as natural hormones would. Rudel *et al.* (2007) have written one of the most exhaustive lists of breast cancer carcinogens and endocrine disruptors, including a substance-specific means of encountering these carcinogens; for example, polycyclic aromatic hydrocarbons (PAHs) can be encountered through first and second-hand tobacco smoke, crude oil runoffs, coal tar residues, vehicle exhausts, industrial combustion byproducts, and waste disposal or waste incineration (Rudel *et al.* 2007).

According to the EPA (2010c), endocrine disrupting chemicals have the ability to mimic natural hormones by causing exaggerated bodily responses (i.e. growth hormones that stimulate increased muscle mass) or by causing bodily responses at inappropriate times. Endocrine disruptors can also interrupt hormone receptors and cause overproduction or underproduction of certain hormones within endocrine glands (EPA 2010c). For instance, Birnbaum and Fenton (2003) suggest that dioxin (TCDD) exposure can be associated with decreased levels of pituitary-released prolactin, and therefore, increased levels of circulating estrogen can be found in the female's blood plasma. These elevated levels of blood-estrogen are considered a prolonging of the 'window of sensitivity to neoplasia', extending the opportunity to develop

cancer (Birnbaum and Fenton 2003). Birnbaum and Fenton (2003) also refer to TCDD dioxin as the most toxic carcinogen ever produced, as it can encourage neoplastic cellular development in practically all histologic sites of the human body.

Agriculture as Environmental Risk

Following World War II (WW2), organochlorines such as PCBs, DDT, toxaphene, heptachlor, heptachlor epoxide, and dieldren, became very bio-available in the United States and globally through agricultural insect control and paper or plastic generating industries (Wolff and Toniolo 1995; Voldner and Li 1995). During the post-WW2 timeframe, concentrations of pesticide and PCB residues in human adipose tissue in the US have shown temporally parallel increase (Wolff and Toliolo 1995). The potential problem is that many organochlorines can mimic estrogens, which means they potentially serve as cancer promoters in the same fashion as steroid hormones (Wolff and Toliolo 1995). Substances that promote estrogen activity within the female body have the potential to cause earlier menses and earlier breast development, both of which associate with higher breast cancer risk (Davis *et al.* 1998). The irony that a dramatic gradient of breast cancer incidence occurred in the US during the 1970s, 1980s and 1990s (SEER 2010b), warrants etiologic investigation of agricultural residues (from organochlorines) in association with elevated breast cancer incidence. Given the developmentally latent nature of breast cancer, it would be scientifically interesting perform a cohort analysis of rurally located girls who were exposed to organochlorines in the 1950s and 1960s. Krieger (1989) considers this type of scenario to be a typical -yet dangerous- situation in which environmentally related breast cancer occurs; mainly because the woman's exogenous carcinogen exposure occurred while her pre-pubescent breast cells were still undifferentiated as a child.

Heavily associated with Illinois agricultural production is a geographic area known as the Lower Illinois River Basin (LIRB), which includes 17,960 square miles of central and western Illinois. This region extends from the downstream end of the 10,950 square miles of the Upper Illinois River Basin (UIRB) at Ottawa, Ill., to the confluence of the Illinois and Mississippi Rivers at Grafton, Illinois (NAWQA 1994). Flowing directly through the LIRB is the Illinois River, a key mode of aqua-transport for Illinois' human, animal, industrial and agricultural wastes (NAWQA 1994). The distinct chemical characteristics of sediment from the UIRB have been identified in sediments of the LIRB (NAWQA 1994). The problem then becomes the potential for river pollutants to enter river alluvia and contaminate aquifers and water supply wells (Groschen *et al.* 2000). During an observation of LIRB monitoring wells and water-supply wells, Groschen *et al.* (2000) observed that six herbicides (atrazine, metolachlor, prometon, bentazon, cyanazine and dicamba) could be readily detected. Of these chemicals, atrazine can be detected in almost 100 percent of water samples from the LIRB. Perhaps this is possible, because approximately 9-million pounds of atrazine were applied to Illinois agricultural fields in 1990 (nearly one-sixth of the national total of applied atrazine), clearly more than any other US State (NAWQA 1994). This represents Illinois' deep reliance upon industrial herbicides. Groschen *et al.* (2000) report that during a 1995-1998 assessment of water quality in the LIRB, dieldrin residues (a prohibited insecticide) were detected in the tissues of every sampled fish and one-third of bed-sediment samples. From this collection of samples, the Sangamon River produced the highest dieldren concentrations reported within the NAWQA's nation-wide water assessment.

Water-resources Investigations Report 99-4229 (Morrow 1999) discussing volatile organic compounds (VOCs) in groundwater of the LIRB identified VOCs in five out of 30

sampled wells, four of them occurring in shallow, glacial-drift aquifers of the Galesburg, Springfield Plain. Morrow (1999) concluded that these substances were the result of both external and internal contamination; internal coming from well water treatment chemicals, and external coming from possible alluvial recharge from the Illinois River. Polycyclic aromatic hydrocarbons such as *benzo(a)pyrene* in the LIRB were detected in concentrations at or above a common reporting level of 50 µg/L in as much as 50 percent of riverbed sediment samples, predominantly near urban areas in locations such as the Vermillion River by Decatur and upstream from Peoria (Groschen *et al.* 2000). River sediments pose a potential threat of entering aquifer alluvia, which would then allow contaminants to reach water wells depending upon aquifer uptake.

Crude Oil Welling as Environmental Risk

Illinois belongs to a vast geologic structure known as the Illinois Basin. The Illinois Basin, an oval depression covering approximately 60,000mi^2 (155,000km^2) within the US Mid-Continent, includes southern Illinois, southwestern Indiana, western Kentucky, far northern Tennessee, and sections of northern and western Illinois (Buschbach and Kolata 1990). The Illinois Basin has produced over 9 billion tons of coal and 4 billion barrels of oil (USGS 1997). Illinois has approximately 650 operating oil fields, including 32,100 active wells (predominantly stripper wells), 12,000 Class II Injection Wells, and 1,700 gas storage sites, the vast majority of which are concentrated in the southern half of the state (IDNR 2010). Since the mid 1980s, oil production in Illinois has declined consistently from production levels of 30,265,000 barrels for the year 1985 (DOE 2010) to just over 15,000,000 barrels of oil for the year 1996 (ISGS 2010a). This suggests that a plentitude of oil wells in Illinois have been either

abandoned or capped due to low oil yields. Currently, some of the biggest oil fields in Illinois are connected to over 2,000 oil wells pulling from ground reservoirs containing over 200,000,000 barrels of oil (ISGS 2010a). These wells and capped sites are another source of environmental risk exposure to carcinogenic substances within the category of aromatic amines and hydrocarbons (Rudel *et al.* 2007). Many of these substances act directly as DNA mutagens and have the ability to interrupt genetic coding. In addition to benzene, polycyclic aromatic hydrocarbons, and other Total Petroleum Hydrocarbons (ATSDR 1999; Gray 2010; Rudel *et al.* 2007), the EPA states that upper Midwestern US soils are more prone to radioactive contamination during the oil extraction process (EPA 2010d), because of the higher geologic tendency for substances such as radium-226, radium-228, and radon, in the upper Midwest. This phenomenon only compounds the already carcinogenic risk of crude oil welling.

Crude Oil Refineries as Environmental Risk

In addition to oil welling sites, Illinois also has four of the 'top-40' US oil refineries (Wood River, Joliet, Robinson, and Lemont refineries), refining up to 973,600 barrels of oil per day (DOE 2010). Unlike the majority of hydrocarbons that are typically hydrophobic (water fearing), a wide variety of liquid aromatic hydrocarbons such as Benzene 1800 PPMV, Toluene 470, Ethyl Benzene 150, and Xylenes 150 (BTEX) are highly soluble in water and are the primary volatile effluents released with desalter waste when refinery tanks are cleared of tank-damaging salt (Worrall and Zuber 2010). Benzene concentrations within desalter effluents can often range from 20 milligrams per liter (mg/L) to 200 mg/L (Worrall and Zuber 2010), which is in great excess of the EPA's MCL of 0.005 mg/L for benzene (Stewart *et al.* 2001). This does not discount the fact that other aromatic substances could still be present in the effluents as

well. In addition to aromatic hydrocarbons, toxic elements such as lead and mercury were noted to be present in refinery effluents (IEPA 2009) of the Wood River Crude Oil Refinery (located in Roxana, IL) discharged from effluent outfalls or stored on-site in ponds and sludge areas (IEPA 2009). One could argue that oil refinery wastewater represents an opportunity for ground water and surface water to interact with toxic substances and known carcinogens.

Hazardous Wastes as Environmental Risk

The EPA classified hazardous waste into four categories, as follows (EPA 2012a):

1) **Listed Wastes**: Hazardous; typically the result of manufacturing, industrial, commercial or chemical byproduct
2) **Characteristic Wastes**: Ignitable, corrosive, reactive, and/or toxic
3) **Universal Wastes**: Batteries, pesticides, mercury and lamps
4) **Mixed Wastes**: Both radioactive and hazardous chemicals

The predominant pathways for exposure to hazardous wastes include inhalation, ingestion and physical skin contact (EPA 2012d). It suffices to say that living within a certain distance of a putative source increases the likelihood of exposure. Often, hazardous waste exposure results from low intensity, chronic exposure to locations such as commercial landfill sites, which was the case in the classic Love Canal scandal where a residential area in Niagara Falls, NY, was exposed to toxic waste dumps (Meade and Emch 2010).

The persisting question regards the pathway of exposure to hazardous waste (Vrijheid 2000). It is possible to be located immediately adjacent to a landfill site, but your drinking

water could be imported from a far away water reservoir. In other words, it is difficult to pinpoint the potential route of exposure when different routes are possible.

Power Plants as Environmental Risk

Over one-half of the United States' electricity is the result of thermal coal plant activity, resulting in a tremendous quantity of coal combustion byproducts such as fly ash, bottom ash and boiler slag (Kalyoncu 1997). Coal processing plants produce large volumes of solid and liquid waste, and the majority of these wastes are stored in on-site impoundments such as wastewater reservoirs or subterranean pits (IEPA 2010; Jüngten and Klein1977; Union of Concerned Scientists 2009). In 1997, approximately 4,600 metric tons of coal ash (including fly and bottom ash) were stored on-site at coal combustion plants in the United States. Illinois, Indiana, Kentucky, Ohio, Wisconsin and Michigan comprised the highest coal ash producing region of the US (Kalyoncu 1997). Given an ever-increasing demand for electricity, the advanced Integrated Gasification Combined Cycle (IGCC) –a method whereby gas circulates through a combusting turbine to generate electricity- has become a central electricity generating technique in the US (Union of Concerned Scientists 2009). However, waste products of coal gasification are concentrated with toxic, refractory organic compounds that must be converted to other substances such as methane and carbon dioxide to minimize toxicity (Khan *et al.* 1981). The issue then becomes discarding residual compounds and metals without contaminating surface or ground water sources.

In 1997, the United States produced 45,480 metric tons of coal ash byproduct, 37,196 metric tons (68 percent) of which were disposed within locations such as wastewater reservoirs, subterranean pits, and abandoned coal mines (Kalyoncu 1997). During this same year, roughly

10 percent of total disposal ash (or 4,600 metric tons) were stored on-site at the same location responsible for combusting the coal (Kalyoncu 1997). This is to say that US power plants and industrial facilities stored 9,800,000 lbs of ash waste on-site during the year of 1997. As of April 2010, there were 24 power plants with a total of 83 ash impoundments in Illinois, 31 of which had low-permeability liners to help prevent leaching into groundwater (IEPA 2010). These 24 power plants were assessed for the likelihood of ash precipitate to recharge into aquifers, and ten (10) of the 24 plants were recommended as high priority (priority 1) areas for aquifer recharge potential, and five (5) of the 24 areas were recommended as medium-high priority (priority 2) areas for aquifer recharge potential (IEPA 2010).

The effluent constituents of coal waste impoundment sites comprise any combination of unusable coal rejects, coal ash residues, and coal coking or coal gasification condensates (Huggins *et al.* 2009). These effluents often contain MCL exceeding levels of metals and inorganic compounds (i.e. mercury, lead, copper, selenium, sulfate, arsenic and ammonia) (Huggins *et al.* 2009), as well as polycyclic aromatic hydrocarbons and other volatile and non-volatile organic compounds (i.e. phenols, benzenes, naphthalenes, and cyanide) (Jüngten and Klein 1977).

At a typical #6 Herrin Coal processing plant in southern Illinois, unusable mineral refuse is rejected from the plant at both the rotary breaker (initial processing) phase and at the coal tailings (post-processing) phase; rejects are then forwarded to highly toxic wastewater impoundment sites such as ponds and artificial bodies of water (Huggins *et al.* 2009). Similar to fly ash impoundment sites, coal wastewater impoundments can threaten aquifers and other water table sources, since effluents can leach and plume from holding areas.

Coal Borings as Environmental Risk

Of the roughly 200 billion tons of coal that are estimated to lie underneath Illinois, an approximate 38 billion tons are considered economically recoverable (ISGS 2010f). Since the mid-19th century, Illinois has hosted 6,649 productive coal borings (see Figure 2.1), including strip, shaft, slope, drift, abandoned, or other uncertain types of coal borings (ISGS 2010b). The majority of Illinois' geolithic rock is saturated with a variety of seams and pockets of bituminous coal (ISGS 2010b; Treworgy and Jacobson 1986). However, not all coal seams are active due to high sulfur concentrations that surpass technological filtration capabilities. Multitudinous elements can be withdrawn from the earth through coal borings, to include arsenic, chlorine, mercury, uranium, thorium, lead, selenium, and more (ISGS 2010c). The same principle applies for inorganic and organic compounds such as sulfate, benzene, naphthalene, phenol, cyanide (ISGS 2010c; Jüngten and Klein 1977; Huggins *et al*. 2009). Given the vast quantity of coal borings (many being surface mines) that have proliferated throughout Illinois, it is easy to assume plethoric chemical exposure from mineral boring activity.

Radium in Illinois Aquifers as Environmental Risk

Alpha and gamma radiation are released during the decay of radium-226, while low-energy gamma radiation and beta particles are released during the decay of radium 228 (EPA-e 2010). Exposure to these energy types increases the risk of cancer (EPA 2010d), just like most other forms of radiation. Throughout a majority of northern Illinois, public water supplies have tested positive for naturally occurring radium concentrations in excess of the EPA's MCL of 5 picocuries per liter (pCi/L) for radium-226 and 228 (USGS 1999).

Deep granite bedrock aquifers in northern Illinois often contain radium 226 and radium 228 concentrations in excess of the EPA's radium MCL of 5pCi/L (IDPH 2008). Similarly, in northeastern Illinois (the most densely populated region of Illinois), naturally occurring radium levels in the Ironton/Galesville, Mount Simon, Ancell (St. Peters group), and Gelena/Platteville sandstone aquifers exceed the EPA standards for radium (Kelly 2008). Some of these wells, such as those sampled in Lake County withdrawing from the Mt. Simon aquifer, produce radium concentrations in excess of 60 pCi/L (USGS 1999), providing high radium exposure risk in drinking water. A potentially offsetting factor to this dilemma is that the majority of water supplies in the Chicago metropolitan district are municipally managed and placed under greater standards for contaminant monitoring, restriction, and filtration.

It is still worthy to note that women who are exposed to radiation for longer periods of time have a higher risk of breast cancer development (Cole and Macmahon 1969). Given the likelihood of women in the Chicago-land area to remain within the Chicago metropolitan statistical area during their lifetimes, it could be argued that they are at a higher lifetime risk of cumulative radium exposure. However, Davis *et al.* (1998) suggest that the timing of exposure to radioactive substances (such as during pre-menarche in early childhood) is more important than the concept of 'cumulative lifetime' exposure to carcinogenic substances.

The Mahomet Aquifer is another location in Illinois that presents waterborne radionuclide risk exposure. During a 1995-1998 assessment of Mahomet Aquifer water quality, all (30) aquifer samples tested positive for radon detection, with concentrations ranging from 110 to 730 pCi/L and a mean concentration of 190 pCi/L (Groschen *et al.* 2000). The EPA MCL for radon in drinking water is 300 pCi/L

Arsenic in Illinois Aquifers as Environmental Risk

As of January 23rd, 2006, the EPA lowered the MCL for arsenic in public water supplies from 50 micrograms per liter (μg/L) to 10 μg/L (EPA 2010b). The USGS (2006) claims that prolonged consumption of drinking water in great excess of arsenic health standards is the most serious arsenic-related health hazard in the United States and throughout the world. Naturally occurring arsenic shares a strong geochemical association with basin-fill deposits of alluvial-lacustrine origin, as well as with mining wastes and landfills (Welch *et al.* 1988; Korte and Fernando 1991). Alluvial and glacial aquifers of the upper Midwest are known to contain high sulfide mineral concentrations associated with glacio-fluvial deposits of ferric-oxide, altogether contributing to the litho-chemical constituency (or higher arsenic concentration) of Midwestern groundwater (Welch *et al.* 2000). Almost 50 percent of the community-supply wells in Illinois are open to this Midwestern aquifer system, exposing people to concentrations of arsenic commonly in excess of the EPA's MCL (Warner *et al.* 2003). This suggests that roughly 50 percent of the state's private well users are at high risk for exposure to toxic levels of arsenic.

Surface and groundwater can experience high arsenic concentrations due to mining of sulfide-bearing rocks (Welch *et al.* 2000); such rocks occur in Illinois at concentrations of 3-5 percent dry weight in many Pennsylvanian rocks (ISGS 2010d). Hard rock mining research (Moore 1994) has shown that the mining of sulfur-bearing rocks can -through watershed transportation- result in higher arsenic concentrations in groundwater and sediment for hundreds of kilometers downstream from mining areas. This suggests that Illinois surface water and groundwater could have a general, topographical vulnerability to arsenic originating from coal mining activities.

In east-central Illinois, the Mahomet aquifer is the primary groundwater resource for public and private water supplies (ISWS 2009b). During an IEPA study of 2,771 community water supply wells from 1978 to 2001, the highest and most frequent [municipally detected] arsenic concentrations were from deep bedrock valley portions of the Mahomet aquifer underlying central Illinois (Warner *et al.* 2003). Many communities, industries, and irrigators depend on the Mahomet Aquifer for their water supply. Withdrawals for irrigation, principally in Mason and Tazewell Counties (also referred to as the Havana Lowlands area), put usage well beyond 100 million gallons per day (Mgal/d) (ISWS-a 2009). One-third, or roughly 71 Mgal/d, of this quantity is utilized municipally (ISWS 2009b), which is significant given that population projections for the Mahomet Aquifer region may increase by 100,000 people to a total of 900,000 by 2020 (ISWS 2009c).

Also, the potential for agricultural chemical and nutrient contamination of groundwater is of concern in the sandy areas of Mason and Tazewell Counties (ISWS 2009c). A hydraulic window connecting the Mahomet aquifer to the Sangamon River allows Mahomet aquifer water to discharge to the river under normal conditions, but allows the Sangamon River to recharge the aquifer when the river is high or when the aquifer is pumped (Mehnert *et al.* 2004). In 1985, the Tazewell County Health Department sampled 590 water wells (municipal and private) for arsenic contamination; 202 of the samples (34 percent) had arsenic concentrations at or above 50 µg/L, and 350 of the samples (59 percent) had arsenic concentrations at or above 10 µg/L (ISWS 2009c). Treatment of these water supplies has been a recommended means to decrease arsenic concentrations.

As mentioned before, arsenic has been found in some wells of the Mahomet Aquifer, approaching or exceeding drinking standards (ISWS 2009c). During a specific assessment of

Mahomet Aquifer constituents from 1996 to 1998, arsenic levels were found in well samples as high as 83 μg/L, with 83 percent of the wells exceeding 1 μg/L and 43 percent of the wells exceeding the EPA's MCL of 10 μg/L for arsenic (Warner 2001). An area northeast of Decatur has been found to contain elevated concentrations of dissolved minerals, possibly as a result of upwelling from the underlying bedrock (ISGS 2010d).

2.4 Spatial and Statistical Analytics of Health Hazards

Retrospective analyses of health outcomes in association with pollution sources commonly rely upon surrogate pollution measures or proxies to account for data not collected during or prior to disease morbidity (Lawson 2006). Lawson (2006) encourages analysts to evaluate a sufficiently large number of pollutant surrogates, in order to reduce unaccounted residual effect in models. All applied inferences drawn from models must regard the ecological effects of spatial scale and data aggregation. These effects can lead analysts to commit ecological fallacies, particularly when modeled significance is extrapolated to different spatial scales and levels of data aggregation (Waller and Gotway 2004). Many limitations pertaining to the analysis of health data stem from the level of aggregation in data, and unfortunately it can be difficult to acquire individual level health data (Lawson 2006). Lower resolution spatial groups, such as census tracts, ZIP codes, or counties are more easily obtainable due to the better upholding of confidentiality concerns.

Classic spatial epidemiologic thought encourages the consideration of outliers in space, because spatial outliers could illustrate locations of confounding effects (Waller and Gotway 2004). One such spatial outlier is the statistically significant cluster of disease incidence.

Clustering is a derivative of spatial autocorrelation that measures how clumped or sparse a spatial feature is with respect to its attribute values (Lee and Wong 2005). The guiding assumption of clustering hypothesis tests is that locations and variable attributes occur independently across space, or that clustering does not occur within the study area (Anselin 1995). In order to calculate Local Indicator of Spatial Autocorrelation (LISA) statistics, it is necessary to quantify a Local Moran's Index (I_i) value for each spatial feature within the study extent. Moran's Indices must be compared with expected Index values and interpreted through their standardized z-scores (Anselin 1995), to determine which Moran's Indices occurred either deliberately or via haphazard chance. The LISA technique was utilized previously by Oyana and Margai (2010) to identify Chicago neighborhoods at high risk for exposure to waterborne lead, helping to identify risk associations within an environmental/demographic framework. Focused clustering tests have also recently been utilized to help researchers confirm or refute increased disease risk around sources of environmental exposure (Guajardo and Oyana 2009; Oyana and Lwebuga-Mukasa 2004). Guajardo and Oyana (2009) and Oyana and Lwebuga-Mukasa (2004) utilized the Lawson Waller Score Test (Lawson 1989) and Bithell's Linear Risk Test (Bithell 1995 and 1999) to illustrate increased risk for breast cancer and asthma, accordingly, in close proximity to putative sources of pollution.

2.5 Exploratory Data Analysis

Enhanced interactivity empowers the map inspector to visualize spatial patterns and formulate richer hypotheses (Adrienko *et al.* 2000). Combinations of graphic visuals can be employed to help facilitate exploratory data analysis. In most instances, exploratory data

analysis is a visual process where the majority of pattern recognition is derived from visual tools such as scatterplots, histograms, and other graphic figures (Edsall *et al.* 2008; Gelman 2004; Tukey 1972). Adrienko and Adrienko (1999) illustrate the process of "making multiple comparisons of a dataset within the map interface." This technique represents a process where mapmakers use scatterplots or other graphic figures to enhance visualization. This can help analysts to uncover spatial patterns; it also assists in the generation of focused research questions. Other investigations have led to competing conclusions about the usefulness or validity of graphic visuals, favoring other graphics such as statistical tables (Gelman 2011). However, it could be argued that statistical tables (like sophisticated legends) push the level of map involvement beyond what is normally conducive to exploratory data analysis. The concept of map involvement was introduced by Alan MacEachren (1982), who focused primarily on the geometric complexities of polygonal data structures. Maps that provide too much information can be cumbersome to interpret and potentially lead audiences toward unintended conclusions (Slocum *et al.* 2009).

2.6 Limitations of the Study

Exposure measurement error is worthy of discussion within an epidemiologic study. Jurek *et al.* (2006) illustrated through an assessment of 57 epidemiological studies conducted by other researchers that 22 of the studies (39 percent) failed to analyze exposure measurement error, and, of the other 35 studies (61 percent) only one of them performed this analysis quantitatively. Jurek *et al.* (2006) recommends (because of the inherent possibility of random and systematic error in analysis) that individuals address exposure measurement error. The

measure of exposure error in this study is complicated by the uncertainty of whether breast cancer incidence in a given ZIP code was the result of exposure to putative risk factor(s) in the observed ZIP code or the result of exposure that occurred in distant locations that are unobserved or uncorrelated by the study.

Medical researchers recognize a latency period where cancerous cells accumulate for periods that extend to multiple decades before cancer detection typically occurs (Nordling 1953; Goldsmith 1987). In order to conduct a retrospective analysis of environmental cancer risk it is important to model exposures as they would have occurred during the cancer latency period. This is perhaps the single largest challenge during a retrospective study, because data are frequently not recorded during the cancer latency period (Lawson 2006). Such instances commonly require surrogate measurements of environmental exposure. In other words, a retrospective analysis is potentially vulnerable to type 1 errors where the null hypothesis of no association is arbitrarily rejected. The clinical detection of cancer occurs many years after the onset of carcinogenesis, inferring that related exposures occur 20 to 40 years before clinical detection. This study is concerned with exposures that could have occurred from the 1940s to the 1980s.

It is notable that all ISGS Well coordinates in this study could in some instances be incorrect by +/- 100 feet, or in rare cases as much as one mile (ISGS 2010e). This would affect the count of wells per ZIP code, predominantly as a result of Wells located near ZIP code boundaries.

The ZIP code scale of disease measurement leaves uncertainty about individual exposures, because the ecological fallacy prohibits areal inference to be extrapolated to the individual. The data in this analysis cannot prove that individuals were exposed to modeled

environmental risk factors or whether the individuals resided in a given ZIP code for a specific length of time. Additionally, chemical information such as soil contents or water constituency was not observed in this research.

Another confounding limitation is the possibility that discrete, non-environmental risk factors played an etiological role in breast cancer incidence. Clinical information and biographical information were not observed in this research. Patient information was restricted to ZIP code case counts provided by the Illinois Department of Public Health. Personal risk factors (genetic or behavioral) can be attributed to breast cancer risk, such as age, inherited genetics, adipose tissue density, one first degree relative with breast cancer history, late age at first full term pregnancy (>30 years of age), early menarche (<12 years of age), late menopause (greater than 55 years of age), never breastfed a child, postmenopausal obesity, alcohol consumption, height (tall), and high socioeconomic status (ACS 2009).

In contrast to human intrinsic etiology, prior research (Lichtenstein *et al.* 2001) has argued (in a cohort analysis of twins) that over 60 percent of cancer outcomes are correlated with discrete, environmental risk factors. The other 40 percent of outcomes were attributed to either discrete, human intrinsic factors or human intrinsic factors associated with environmental risk factors (Lichtenstein *et al.* 2000). There are a vast number of ways that cancer etiology can be considered (Krieger 1989). It is sensible to consider cancer as a complex epidemiologic variable. Through spatial analysis, biostatistics, and GIS, it is possible to evaluate cancer outcomes in association with environmental risk factors and extend our understanding of environmental breast cancer risk.

CHAPTER 3

METHODOLOGY

3.1 Study Design

This research is a retrospective case/control analysis of Illinois breast cancer incidence during a study window of 1996 to 2000. The study is designed to evaluate the spatial and statistical associations between potential sources of environmental pollution and breast cancer incidence at the Illinois ZIP code scale. Breast cancer health data for this study reflect the time period of 1996 to 2000. The purpose of this study is not to prove causation between breast cancer incidence and environmental risk factors, but rather the purpose is to illustrate associations between breast cancer incidence and suspected environmental risk factors (Figure 3.1 illustrates these factors).

3.2 Software and Data for Study

Software

The software utilized to conduct this research included Microsoft Office *Excel 2010* ™, Microsoft *Excel Poptools*™ Add-in, Environmental Science Research Institute *ArcGIS 10* ™, Biomedware *SpaceStat 3.5* ™, Biomedware *ClusterSeer 2.3* ™, and SAS Institute *SAS 9.2* ™.

Breast Cancer Data

Breast cancer count data were obtained from the Illinois Department of Public Health (IDPH), representing breast cancer cases that were recorded at the ZIP code scale during the study period of 1996 to 2000. Breast cancer counts represented this five year window. Individual annual counts were not provided. Breast cancer data were provided in three specific age groups: 18 to 44, 45 to 64, and 65 and over.

Environmental Risk Factors evaluated in the Study
(n=sample size, Mu=mean frequency per ZIP code)

From the US Energy Information Administration, USEIA:
Power Generation Facilities: coal, coal/gas & nuclear (n=29; Mu=0.02)

From the Illinois State Geologic Survey, ISGS:
Coal Borings: slope, shaft, drift, strip, methane & test pit (n=30,992; Mu=23.07)
Oil/Gas Exploration & Production Wells (n=47,785; Mu=35.58)
Oil/Gas Injection Wells (n=923; Mu=0.68)
Oil/Gas Storage & Observation Sites (n=905; Mu=0.67)
Plugged & Abandoned Wells (n=51,349; Mu=38.23) †
Water Wells (n=279,885; Mu=208.40)

From the Environmental Protection Agency, EPA:
Air Release Facilities, AFS (n=675, Mu=0.50) †
Large Quantity Hazardous Waste Generators, RCRA (n=787, Mu=0.56)
Pesticide Producing Facilities, SSTS (n=733, Mu=0.54)
Superfund Sites, CERC (n=49; Mu=0.03) †
Toxic Release Inventories, TRI (n=2829; Mu=2.10) †
Wastewater Releases, NPDES (n=256; Mu=0.19)

From the National Agricultural Statistics Survey, NASS:
Percent Corn and Soybean Agricultural Land Cover (n=all ZIP codes, Mu=0.48)

† Variable dropped from the model due to detected collinear effects (rho > 0.60)

Figure 3.1. Total list of environmental risk factors (and sources) considered in this study.

Population Data

ZIP code population data were acquired from the US Census Bureau's data download center (US Census Bureau 2010d), reflecting Illinois ZIP code populations from the US 2000 Census. The sampling space for this research included 1343 Illinois ZIP codes and a total of 4,803,505 females who were 18 years of eight and older.

Study Area Feature Class

A polygon feature class of Illinois ZIP code tabulation areas (ZCTAs) for the year 2000 was acquired through the US Census Bureau (US Census Bureau 2010c), enabling the spatial evaluation of data in a GIS.

Environmental Risk Factors

The independent variables analyzed in this study included georeferenced sources of environmental risk recorded by the Illinois State Geologic Survey (ISGS 2010e); the US Department of Agriculture's National Agricultural Statistics Service (USDA-NASS 2012); the EPA's Geospatial Data Entry Project (EPA 2012b), the US Energy Information Administration (USEIA 2009), Nuclear Energy Information Service (NEIS 2010), and the Illinois Environmental Protection Agency (IEPA 2009).

The ISGS' wells and borings database includes longitude and latitude point coordinates for over 556,541 observations, including recorded types such as plugged and abandoned wells, oil and gas producing wells, oil injection wells, mineral borings, water wells, and gas storage/observation wells. Data from the EPA Geospatial Data Entry Project (for the state of Illinois) includes 5,800 point coordinates of potential pollution sources such as wastewater

releases (NPDES), air pollution facilities (AFS), large quantity hazardous waste generators (RCRA), toxic release inventories (TRI), and pesticide producing facilities (SSTS). Data from the USEIA, NEIS, and IEPA, provide locational information for major power generation plants in Illinois. The types of power plants selected for this study included nuclear, coal, and gas fueled power plants.

Remotely sensed land cover classification data for Illinois were classified by United States Department of Agriculture-NASS using LandSat 5 TM and LandSat 7 ETM+ images from 1999 and 2000 (USDA-NASS 2012). The spatial resolution was 30 meters x 30 meters, making this imagery appropriate for GIS analysis. Classification accuracy of the aerial imagery was rated at 85 percent to 95 percent for agricultural classes (i.e. corn, soybean, and wheat). For this research, corn and soybean pixel counts were extracted from the USDA-NASS (2012) aerial mosaic using Zonal Statistics in ArcMap and summarized as "percentage Corn and Soybean Land Cover" within ZIP codes.

Modeling of Risk Factors

Due to the prevalence of crop rotation cycles, Corn and Soybean Land Cover proportions were summed into a single land cover proportion per ZIP code, representing both crop types. The modeling of Corn and Soybean Land Cover was employed as a surrogate measure for exposure to pesticides, herbicides, and fertilizers. It was further assumed that the intensity of agricultural practice would associate with an increased likelihood of exposure to agricultural chemicals.

Exposure to Water Wells was modeled as the number of "wells per 'at risk' female". To calculate this, the frequency of Water Wells within any given ZIP code was divided by the population of 'at risk' females.

The maximum frequency of Power Plants per ZIP code was 1 (one); therefore, power plants were modeled in absolute frequency per ZIP code. Essentially, the Power Plant effect was either "on" or "off" given the binary scale of Power Plant frequency.

All other risk factors (Mineral Borings, Oil/Gas Wells, Oil/Gas Storage Observation Sites, Oil/Gas Injection Wells, Hazardous Waste Generators, Air Release Facilities, Pesticide Producing Facilities, Wastewater Releases, Superfund Sites, and Toxic Release Inventories) were modeled as *a product of* 'at risk' population density per ZIP code. In these exposures, risk factor frequencies were multiplied by the population density of the 'at risk' population in each ZIP code. The justification for this approach was two-fold. First, the ZIP code scale of the disease data promoted the modeling of areal-level exposure (and population density is area dependent). Second, it was assumed that exposure likelihood would increase as the frequency of risk factors increased in relation to population density. This helped to account for population dynamics that commonly distinguish rural populations from urban populations, and vice versa. See Appendix A for a cartographic library of the spatial distributions of environmental risk exposures as they were modeled within this research.

Retrospective Modeling of Risk Factors

To a very large degree, the majority of environmental risk factors in the study have maintained a consistent presence during the 20th century or earlier. Coal mining has been a consistently prevalent activity since the 19th century (ISGS 2010b), just as Oil and Gas Welling

has been a common place Illinois industry since the 19th century (USGS 1997). Peak oil production occurred in Illinois from 1955 to 1963 (IDNR 2010), suggesting that a large proportion of the 'at risk' population was exposed to peak oil exploration and production during the earlier stages of the latency window. The vastness of production wells and dry/abandoned wells is a testament to oil welling proliferation in the southeastern Illinois Basin. It is assumed that the majority of Power Plants have been in operation since at least the 1950s, given the long history of coal power generation. Furthermore, Illinois' Nuclear Power plants have been in operation since the 1960s, with two-thirds of them commissioning in the 1980s. Residential use of private Water Wells is another historic phenomenon in Illinois, in which rural dwellers utilize well water for consumption (ISWS 2009a). It is assumed that the intensity of Water Well usage during the cancer latency window is reflective of current data on Water Wells in Illinois. Oil Injection Wells came to popularity in the United States during the 1930s, and their application was not federally regulated until the 1970s (EPA 2012c). Given the lower levels of oil productivity in Illinois, it was sensible to argue that oil injection welling has been used to enhance productivity. This helps to explain the approximate 12,000 Class II injection wells that are utilized in Illinois (IDNR 2010). Corn and Soybean Land Cover was optimized in Illinois by the 1950s and has remained steady (Ramankutty and Foley 1999). As such, the aerial imagery used to model Corn and Soybean Land Cover is likely representative of Illinois' agricultural surface during the cancer latency period.

See Table 4.2 for a partial list of EPA registered companies (not all are shown) that were operational in high risk zones during the latency period. The formation dates of these organizations were screened in order to filter out organizations that started after the 1980s.

3.3 Data Processing and Preparation

Preparing Breast Cancer Data

Crude breast cancer incidence rates were calculated with breast cancer count data from the Illinois Department of Public Health. Crude incidence can be calculated by dividing the number of disease cases that occur within a set population over a given period of time (Rothman 1998). Given that female breast cancer is 130 times more frequent than male breast cancer (Greenlee *et al.* 2000), this study only modeled breast cancer incidence for females. Furthermore, breast cancer incidence was calculated in relation to the 'at risk' female population within ZIP code areas. The 'at risk' population was treated as females age 18 and older.

The formula used to calculate crude breast cancer incidence per 100,000 'at risk' females within any given ZIP code was as follows:

$$\text{Incidence} = \left[\frac{5\ year\ count}{5\ years} \text{ x } \frac{100{,}000}{ZIP\ code\ "at\ risk"\ population} \right]$$

In order to reduce the potentially confounding effects of unequal age distributions among age groups (NYDPH 2006), crude incidence data were adjusted to the US 2000 standard population. Age adjusted incidence was calculated utilizing census data from the 2000 US Census. The following age-standardization formula was utilized as recommended by Surveillance Epidemiology and End Results (SEER 2011):

$$\text{AA Rate} = \sum_{i=x}^{y} \left[\left(\frac{\text{count } ith}{\text{pop } ith} \right)\ x\ 100{,}000\ x\ \left(\frac{\text{standard million } ith}{\sum_{j=x}^{y} \text{standard million } jth} \right) \right]$$

95 percent confidence intervals for incidence were calculated for each age group using the formula outlined by the New York Department of Public Health (NYDPH 2006):

$$95\ \%\ CI = +/- 1.96 \times \text{standard error},$$
$$= +/- 1.96 \times \text{rate} / \sqrt{1343\ ZIP\ codes}$$

Surrogate Control for Potential Frailty and Resistance

In an effort to control for what Lawson (2006) referred to as genetic frailty (a potentially confounding effect), a correlation analysis between US ancestry groups and associated ZIP code breast cancer incidence was conducted. Most researchers utilize gene foci information, inheritance data, clinical data, and otherwise cohort information (Kruglyak *et al.* 1996, Lander and Green 1987, Locatelli *et al.* 2004, Struewing *et al.* 1997) to derive frailty models, but this research accounted for frailty by using Census ancestry data, randomization methods, and Local Indicators of Spatial Autocorrelation. The sample size within this study is essentially deterministic of the Illinois population, allowing observations to be associated with a large number of females, ancestry groups, and breast cancer outcomes. This enabled the evocation of central limit theorem arguments.

First, ZIP code breast cancer incidence was $\log_{10}$ transformed to reduce the degree of over-dispersion. Using Microsoft *Excel Poptools*, a global mean was declared by using Bootstrap reshuffling (with replacement) and Monte Carlo simulations (all Monte Carlo simulations involved 4,999 iterated reshufflings). Bootstrap reshuffling and Monte Carlo simulations were also used to declare a most likely standard deviation. The Z-scores for incidence were calculated using the following equation:

Z-score Incidence = [($\log_{10}$ INCIDENCE) – (Bootstrapped MEAN)] / (Bootstrapped STANDARD DEVIATION)

Z-score Incidence was then correlated with ZIP code Ancestry Group Percentages (i.e. percent Greek, percent Sub-Saharan, etc) reported by the US Census. Table 3.1 illustrates these correlated associations. The Bootstrap Reshuffling and Monte Carlo simulation method was then used to calculate the mean and standard deviation of correlations between Z-score Incidence and Ancestry Group Percentages. These correlations were then converted to Z-scores. Z-scores of the correlations were intended to represent the standardized association between ancestry and breast cancer risk.

ZIP code ancestry percentages were then multiplied by the Z-scores of risk correlations, yielding Ancestry Risk Betas for each ancestry group within ZIP codes. Ancestry Risk Betas were then summated within each ZIP code, yielding Ancestry Sum Betas. Ancestry Sum Betas represented the total level of ancestry-related risk for a given ZIP code. Ancestry Sum Betas were then converted to Ancestry Sum Beta Z-scores, using Bootstrap reshuffling and Monte Carlo simulations.

Next, a Local Moran's *I* procedure was utilized to evaluate Ancestry Sum Beta Z-scores as the input attribute, to obtain the Moran Index values representing neighborhood levels of ancestry-related risk. After obtaining these Moran Index values, a Bivariate Local Moran's *I* procedure was utilized to evaluate the previously acquired Moran Index values as the ego attribute and the Z-standardized Log_{10} Age-Adjusted Incidence as the neighborhood attribute. Clusters were identified with $p<0.05$ significance, using Simes correction. Locations with a high ego value and high clustered neighborhood were considered locations of ancestral frailty,

and locations with a low ego value and low clustered neighborhood were considered locations of ancestral resistance. Results of the Bivariate Local Moran's *I* can be seen in Figure 3.2.

Lastly, the regression coefficient between the Moran Index Values and Age-Adjusted Incidence was observed in a Generalized Linear Model to analyze the significance of the effect of Moran Index Values on Age-Adjusted Incidence. Moran Index Values had a highly significant effect on Age-Adjusted Incidence ($F_{1, 1342} = 6.72$, $p<0.0096$), with an estimated parameter coefficient of 6.7056. Breast cancer incidence was then scaled by using this parameter coefficient to decrease incidence in ZIP codes identified as High-High and Low-Low members in the Bivariate Local Moran's *I*. Incidence was increased in ZIP codes that were identified as Low-Low members and decreased in ZIP codes identified as High-High members, using the following formulae:

[(Moran's Index Value * 6.7056) - Age-Adjusted Incidence]

[(Moran's Index Value * 6.7056) + Age-Adjusted Incidence]

The algorithmic process for scaling incidence by associated ancestral risk can be observed in Figure 3.3.

Table 3.1. Correlation between ancestry group and Z-score Log_{10} Incidence.

Ancestry Group	Correlation with Incidence	Ancestry Group	Correlation with Incidence
Arab	0.11	Polish	0.13
Czech	0.11	Portuguese	-0.01
Danish	0.05	Russian	0.13
Dutch	-0.02	Scotch	0.05
English	0.02	Scottish	0.08
French	-0.03	Slovak	0.08
French Canadian	0.07	Subsaharan	0.12
German	0.05	Swedish	0.02
Greek	0.17	Swiss	0.03
Hungarian	0.18	Ukrainian	0.10
Irish	0.15	United States	-0.07
Italian	0.08	Welsh	0.03
Lithuanian	0.13	West Indian	0.10
Norwegian	0.01		

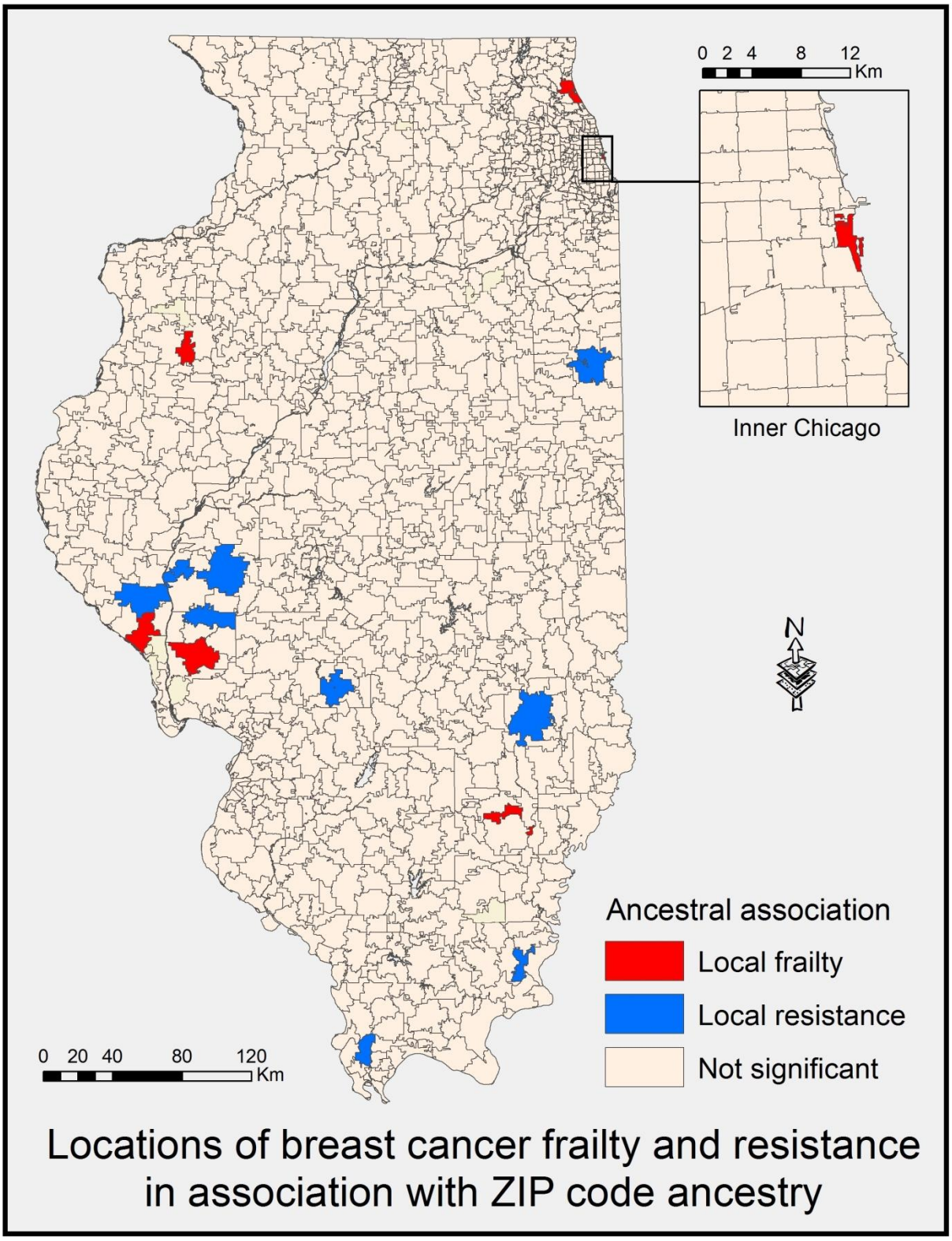


Figure 3.2. Bivariate Local Moran's *I* evaluating the spatial clustering of age-adjusted breast cancer incidence and Census-derived ancestry risk.

Ancestry Frailty/Resistance Algorithmic Flow Chart			
1	Log10 transform of incidence = "**A**"	Bootstrap & Monte Carlo: Mean, Std. Dev. **A**	Z-scores of **A = "B"**
2	Extract ZIP code ancestry percentages from US Census	Correlate ancestry percentages with **B ="C"**	
	Bootstrap & Monte Carlo: Mean, Std. Dev. of **C**	Z-scores of **C**= "**D"**	
3	Multiply **D** by ZIP code ancestry percentages to get Ancestry Betas = "**E"**	∑**E** = Sum Ancestry Betas = "**K"**	Bootstrap & Monte Carlo: Mean, Std. Dev of **K**
	Z-scores of **K** = "**M"**		
4	Implement Local Moran's *I* with **M** as the input attribute	Extract the Moran Index values = "**R"**	
5	Implement Bivariate Local Moran's *I* , with **R** as ego ZIP code and **A** as the neighboring ZIP code		
6	GLM of Age-adjusted incidence on **R:** Coefficient estimate = "**Z"**	If frailty is High-High clustering, then: Incidence - (**R*****Z**)	If frailty is Low-Low clustering, then: Incidence + (**R*****Z**)

Figure 3.3. Algorithmic flow chart detailing the ancestral frailty/resistance measurement.

3.4 Statistical Analysis

Hypothesis Testing and Case/Control Study Space Development

According to Luc Anselin (1993), exploratory spatial data analysis -when performed in conjunction with GIS techniques- should focus on "measuring and displaying local patterns of

spatial association." This encourages the analyst to look for spatial clustering. It is already known that gradients of breast cancer incidence occurred at the ZIP code level in Illinois during the study period (Wang 2004).

In order to challenge the null hypothesis of no spatial clustering, the spatial statistics tool Getis Ord Gi* (Getis and Ord 1995) was used in ArcGIS 10 ™. Polygon contiguity was used to conceptualize spatial relationships within the testing space. All of the 1343 ZIP codes were entered into the clustering test. The overall objective was to identify statistically significant clusters of breast cancer with GiZ Scores greater than or equal to 1.96, and with p-values less than or equal to 0.05. The meaningfulness of spatial autocorrelation hypothesis testing is that locations of clustering reject the null assumption of Poisson or homogenous spatial variability (Waller and Gotway 2004). Spatial analysts suggest that locations of clustering can be influenced by confounding factors such as environmental pollutants or genetic frailty (Lawson 2006; Waller and Gotway 2004).

Clustering analysis identified 57 ZIP codes belonging to spatial clusters. Cluster ZIP codes were marked as binary ones (1's) for evaluation in the case/control model. All non-cluster ZIP codes that were below the global mean (Mu = 126.53 cases per 100,000 'at risk' females) were marked as binary zeros (0's) for evaluation in the model. All other ZIP codes (e.g. non-cluster ZIP codes greater than the global mean) were withdrawn from the model. In sum, the case/control model contained 57 case ZIP codes and 674 reference ZIP codes. The developmental goal was to isolate breast cancer risk zones from non-risk zones, so that environmental risk factors could be evaluated in terms of association with risk and non-risk.

Generalized Linear Mixed Model (GLMM) for Binary Responses

Generalized linear mixed models (GLMMs) for binary outcomes are somewhat similar to logistic regression models, with regard to predicting dichotomous responses to continuous or categorical independent variables (Hosmer and Lemeshow 2000). GLMMs allow parameter estimation within a restricted estimate maximum likelihood (REML) framework, reducing estimation bias of variance/covariance constructs (Meza *et al.* 2007). GLMMs enable the analyst to evaluate dependent responses to a combination of fixed and random effects. The Proc GLIMMIX program was utilized in SAS 9.2 to test whether risk factors expressed a statistically significant outcome with Case or Control groups. Proc GLIMMIX is unique in that it allows the modeling of a random error component that can represent the spatial or temporal covariance within the model (SAS 2012a). As such, a random statement was utilized to account for the spatial structure of residuals, and an exponential type of spatial covariance was selected.

The objective was to test the null hypothesis of no association between independent variables and Case/Control outcomes. If a statistically significant effect was detected between the independent variable and the Case/Control model, then an alternative hypothesis of empirical slope association ($\beta_i > 0 < \beta_i$) would have been supported.

Discriminant Function Analysis

Lewicki and Hill (2007) utilized statistical methods, such as analysis of variance and mean difference, to exercise what is known as discriminant function analysis. Discriminant function analysis allows a researcher to analyze if a variable discriminates between two or more groups by calculating a variable known as the *standardized function coefficient* (Lewicki and Hill 2007). In the discriminant function analysis, the reverse process of a Multivariate Analysis

of Variance (MANOVA) takes place, in which the predictor variables are used to estimate the grouping of observed outcomes. Standardized beta coefficients are given for each variable in each discriminant function, and the larger the standardized coefficient, the greater is the contribution of the respective variable to the discrimination between groups (Poulsen and French 2012). For this research, five predictor variables were observed: Percentage Black Population, Percentage Hispanic Population, Percentage Asian Population, Percentage Without a High School Diploma, and Percentage Employed within Managerial Labor. The variable of Managerial Labor was included in this assessment as a surrogate for financial income. Since Median Household Income was used in the stratification process of random sampling, another variable was chosen in order to prevent redundancy in variable effects. Furthermore, given the high prevalence of Managerial Labor and the normalcy of its distribution, it was a strong variable for the discriminant function. Percent Black Population, Percent Hispanic Population, Percent Asian Population, and Without High School Diploma were successfully transformed to meet the discriminant function analysis assumptions as outlined by Lewicki and Hill (2007), such as low-correlation and lack of influential outliers. Many other socioeconomic predictor variables were considered for the analysis, but they ultimately failed to meet assumptions, even after transformation.

The discriminant function analysis included five case groups comprising 57 case ZIP codes (Figure 3.4) and one reference group comprising 160 reference ZIP codes. The reference group was selected by conducting a stratified random sampling of the original 674 reference ZIP codes. References were stratified into four quantiles according to Median Household Income. Median Household Income was used for stratification because Median Household Income separated risk zones into somewhat discrete areas. (Figure 3.5; counties were used in

this figure to assist with visualization). The first ZIP code within each stratum was randomly selected using a random number generator. From each of the four strata, 40 ZIP codes were randomly selected (without replacement), yielding a total of 160 randomly sampled reference ZIP codes. The scatterplot matrix of predictor variable correlations can be observed in Figure 3.6, revealing that collinearity was absent from the model.

Given that risk zone sizes varied between 15 and 6 ZIP codes per zone, it was decided to evaluate the 'within group variances' instead of 'pooled variances'. The pooled variance criterion was set to false, forcing the discriminant function analysis to assume unequal variances among groups (due to unequal zone population sizes).

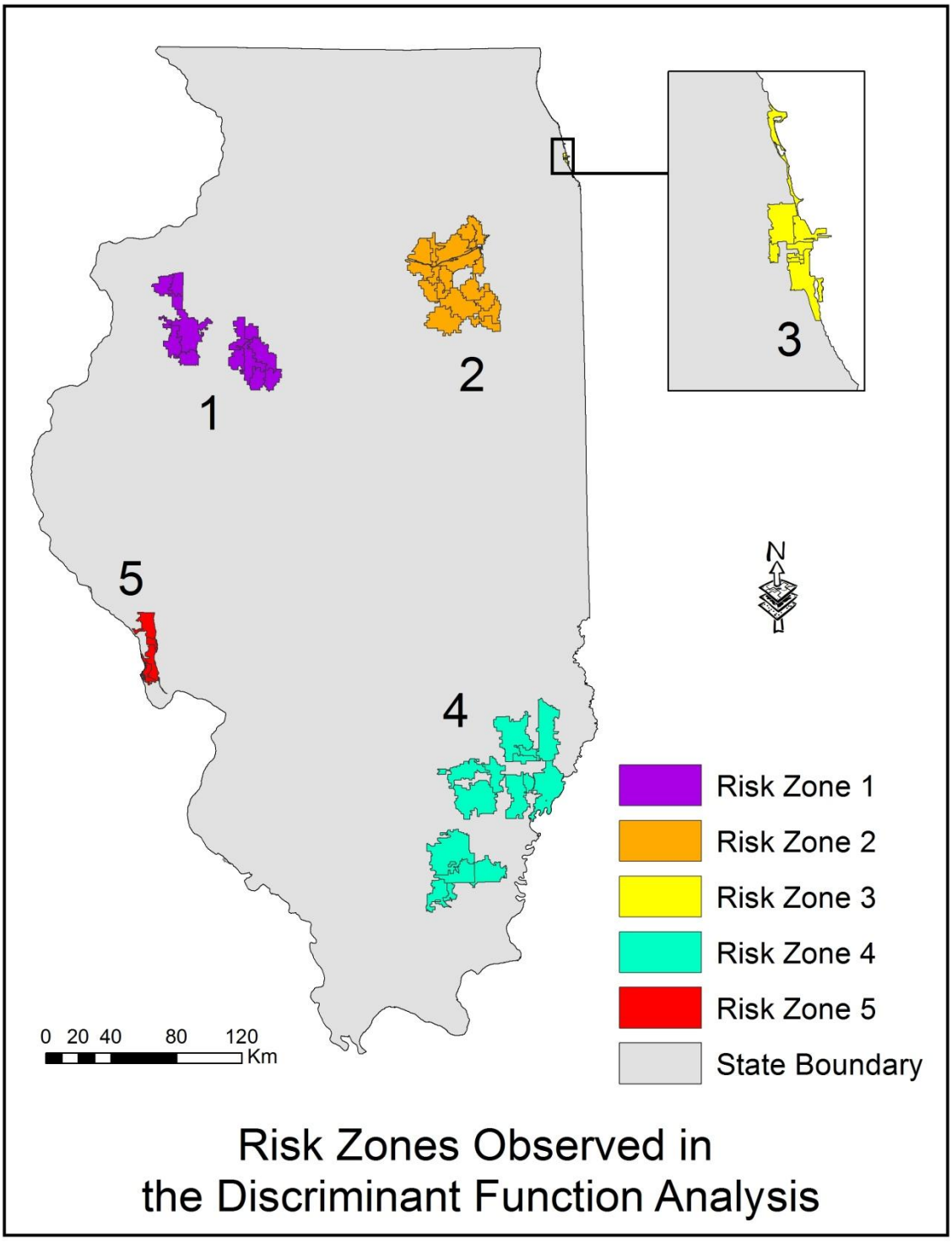


Figure 3.4. Five risk zones observed in the discriminant function analysis. Risk areas represent the locations of statistically significant ($p<=0.05$) breast cancer incidence clustering.

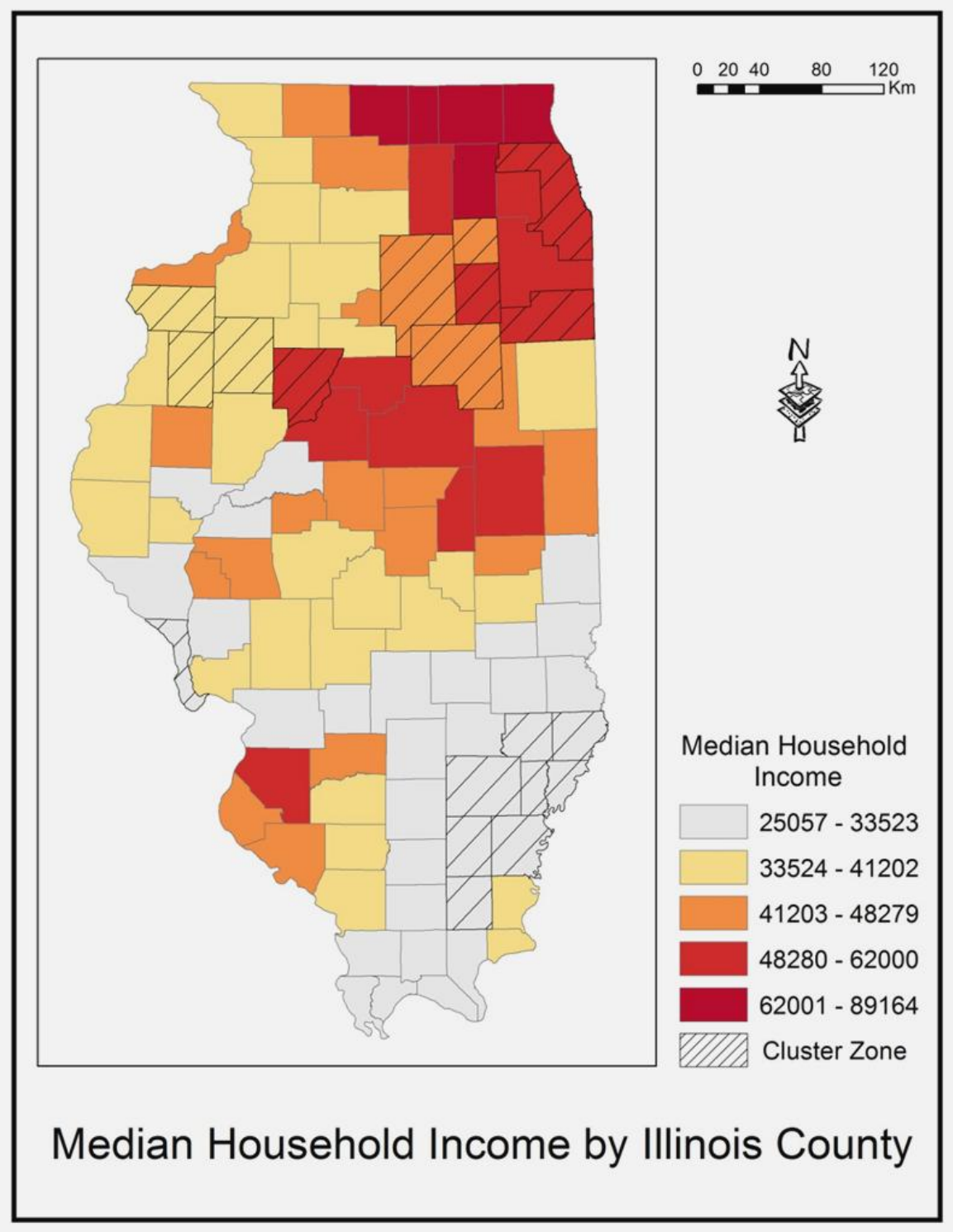


Figure 3.5. Median household income in Illinois and breast cancer risk zones. Notice that the stratification of median household income separates each of the risk zones into distinct areas.

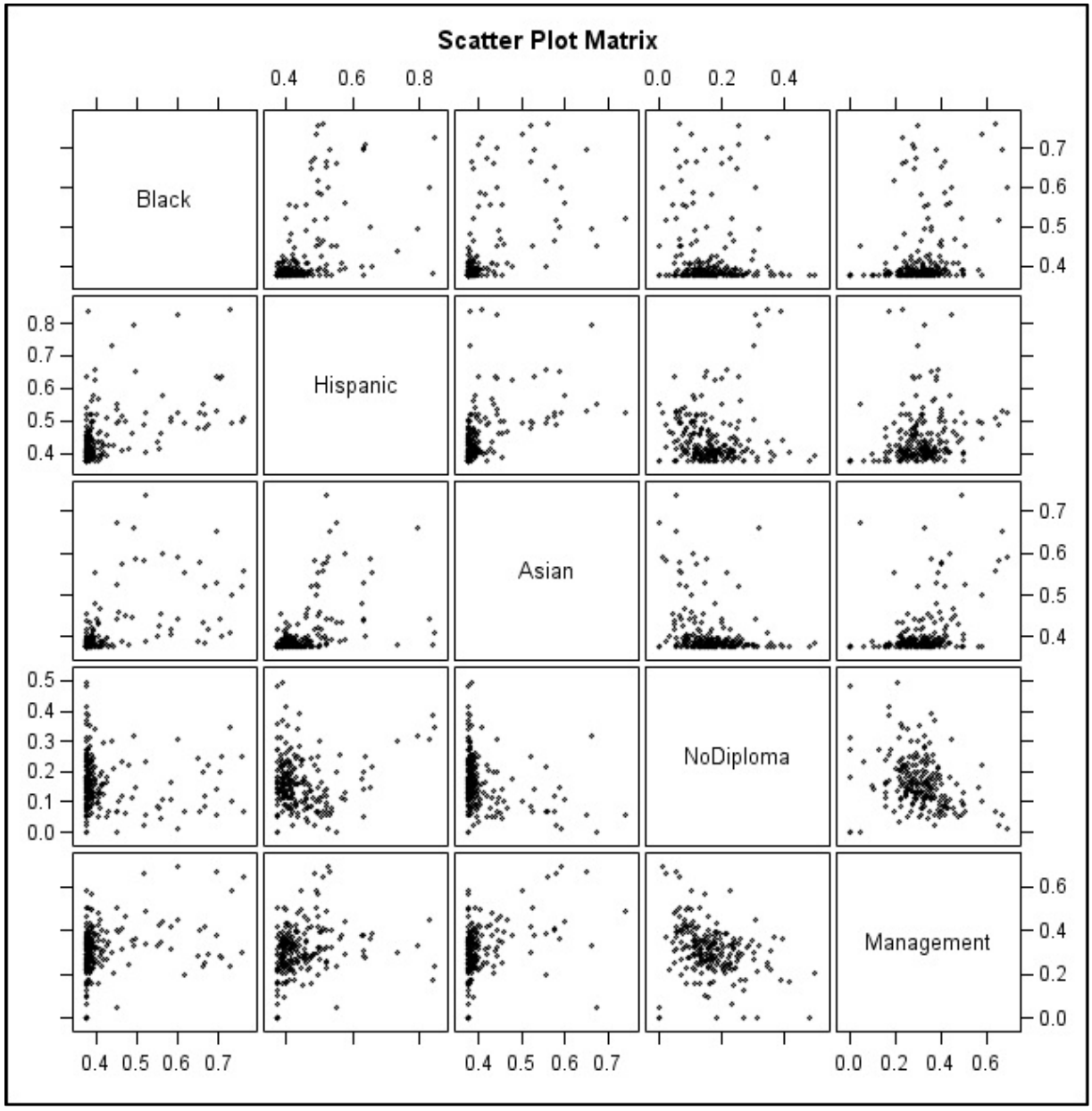


Figure 3.6. Scatterplot matrix showing correlations between predictor variables used in the Discriminant Function Analysis; produced with SAS 9.2.

3.5 Spatial Analysis

Focused Clustering Tests

The Lawson and Waller Score Test and the Bithell's Linear Risk Score Test were used to test the distance decay behavior of breast cancer incidence in spatial relation to focused point sources of potential contamination. Both tests were performed via Biomedware's *ClusterSeer 2.3* ™ software. Locations were selected for these tests according to a 20 and 30 mile radial bandwidth placed around the center of potential contamination.

Lawson Waller Score Test

Lawson (1989) and Waller *et al.* (1992) implemented a scoring method that measures the intensity of disease frequency around a central location of exposure risk. The test statistic for the Lawson Waller Score Test is calculated as follows:

$$T_i = \sum_{i=0}^{n} W_{ij} \left(c_j - n_j \frac{C}{n}\right),$$

where *c* is the number of cases occurring in each at risk population, *n*. *W* is a weight that expresses an inverse distance effect, where disease risk is weighted less as distance between the 'at risk' population and the focus point increases.

Bithell's Linear Risk Score Test

Bithell (1995 and 1999) implemented a focused cluster detection method that was designed to sense excess disease risk near nuclear plants in the U.K. The test results are dependent upon a relative risk function (RRF) that measures linearity in relative risk in relation

to the distance between population and point source of risk. The RRF implemented in this research is as follows:

$$f(d)=1 + \beta/(1+d/\varphi),$$

where *d* is the distance between the focus point and the 'at risk' population, φ (phi) is the rate of decay of cases with distance to the source, and (1 + β (beta)) represents the ratio of risk at the focus normalized by an infinitely far ratio of risk.

CHAPTER 4

RESULTS

4.1 Breast Cancer Incidence in the Study Area

Table 4.1 illustrates a summary of breast cancer incidence by age distribution, including crude incidence, age-adjusted incidence, and empirically Bayesian filtered incidence. As expected, the rates of breast cancer increased as age increased. The degree of over-dispersion decreases across crude, age-adjusted, and empirical Bayesian incidence. It appeared that the age-adjusted and the empirical Bayesian datasets displayed the most Poisson random variation; however, an excessively high frequency of zero counts acted to undermine this initial assumption of Poisson randomness (Lawson 2006) within the crude and age-adjusted distributions.

The US age-adjustment served different useful functions. The initial function was to control for age confounding affects resulting from unstable age distributions across space (Figure 4.1). The second utility of the age-adjustment was to account for the 'old-age population' effect in which cancer risk was elevated in ZIP codes with higher proportions of older females (Figure 4.2). In this instance, the proportion of females age 65 and older had a highly significant effect on breast cancer incidence ($F_{1342, 1} = 27.18$, $p<0.0001$). The linear trend in Figure 4.3 illustrates the flattened old-age effect that results from the age-adjustment to the US standard population.

Table 4.1. Female breast cancer incidence by age distribution. †

Crude incidence				
Age Group	*Mean Rate*	*95% c.i., +/-*	*Variance*	*Zero Counts*
18 to 44	42.71	2.28	12463	608
45 to 64	237.18	12.69	70636	312
65 and Over	436.90	23.37	298682	273
18 and Over	182.07	9.74	20234	183
US age-adjusted incidence				
Age Group	*Mean Rate*	*95% c.i., +/-*	*Variance*	*Zero Counts*
18 to 44	18.68	1.00	2378	608
45 to 64	52.69	2.82	3485	312
65 and Over	55.21	2.95	4771	273
18 and Over	126.57	6.77	12729	183
Empirical Bayesian filtered US age-adjusted incidence				
Age Group	*Mean Rate*	*95% c.i., +/-*	*Variance*	*Zero Counts*
18 to 44	18.24	0.98	56	29
45 to 64	54.86	2.93	160	5
65 and Over	55.03	2.94	136	0
18 and Over	129.31	6.92	780	0

† Rates are normalized per 100,000 females within the age group.

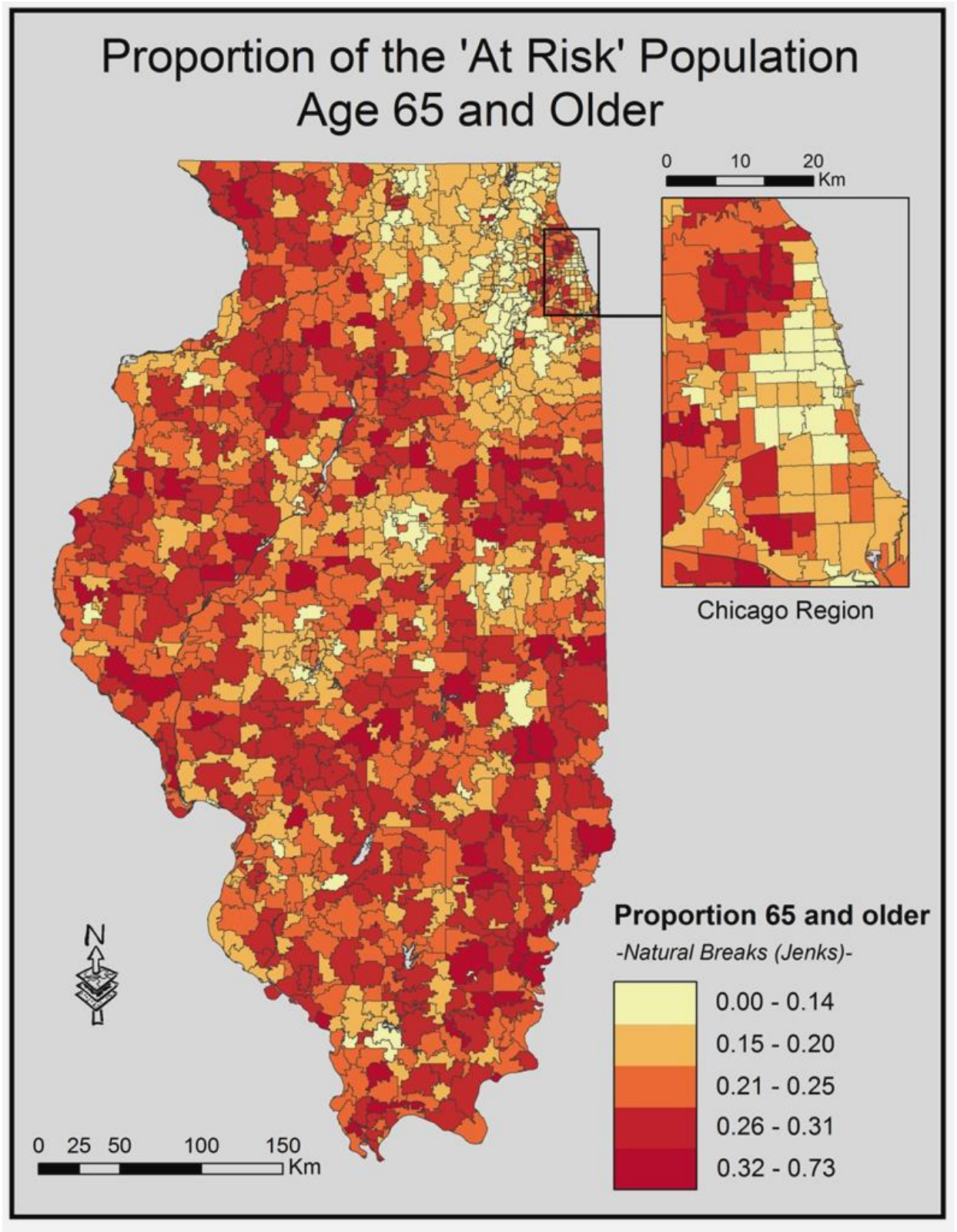


Figure 4.1. Spatial distribution of female population proportions (age 65 and older).

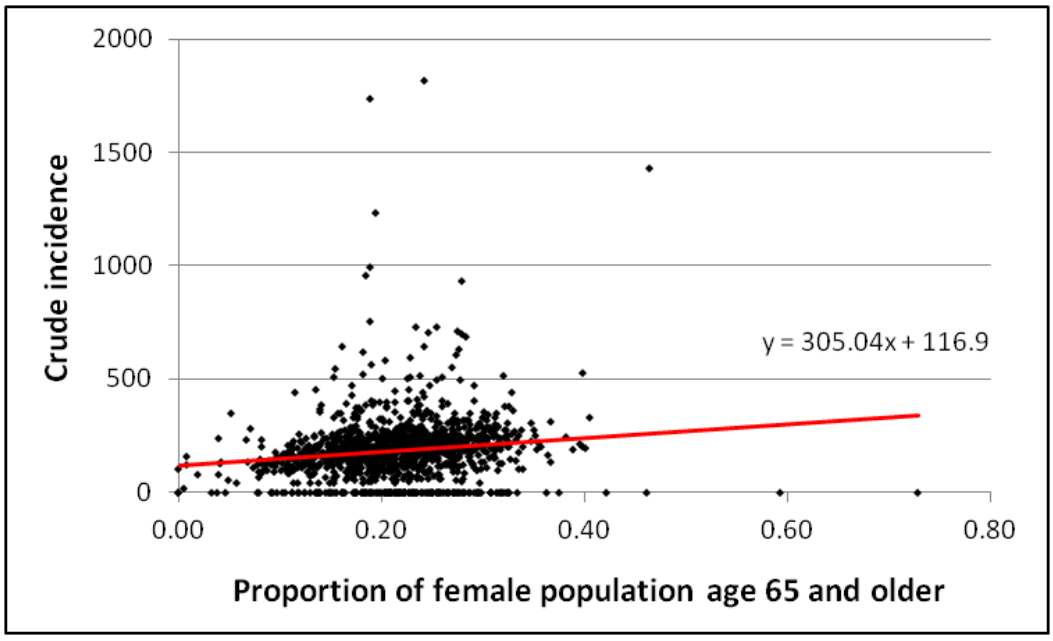


Figure 4.2. Relationship between the proportion of female population (age 65 and older) and Crude breast cancer incidence. The linear trend expresses increased risk due to older age effects, helping to justify the SEER age-adjustment.

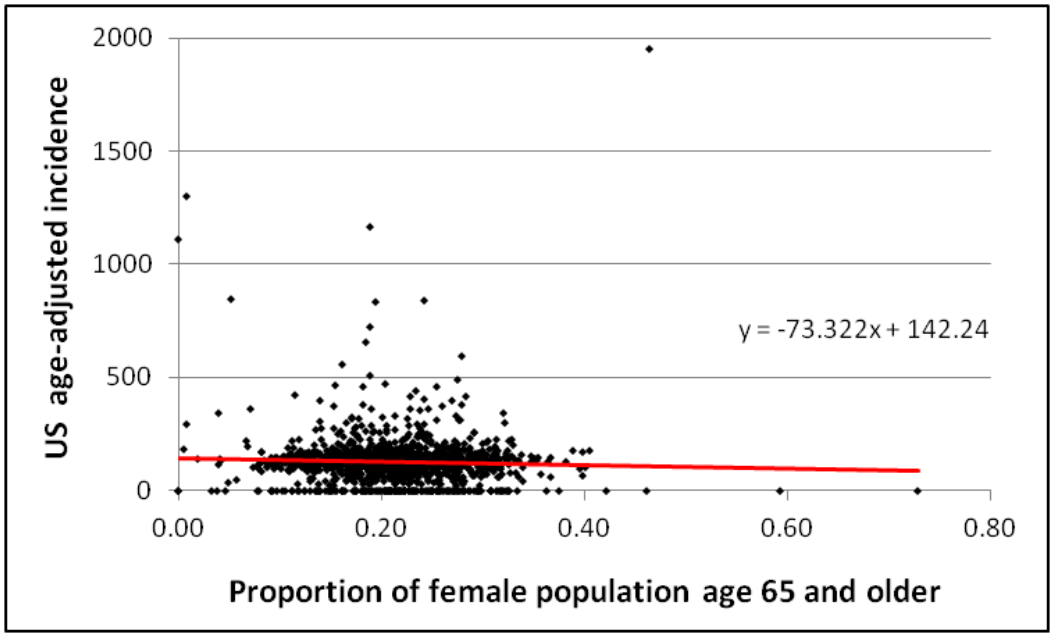


Figure 4.3. Relationship between the proportion of female population (age 65 and older) and US age-adjusted breast cancer incidence. The linear trend illustrates the removal of risk associated with older age effects. Notice that the slope is flatter than the Crude incidence slope.

4.2 Exploratory Data Analysis

After observing the linear association between ZIP code female population and ZIP code breast cancer counts (Figure 4.4), it was apparent that the count range was more unstable as female population increased. The linear regression of female population size on breast cancer counts displayed a highly significant effect ($F_{1341, 1} = 9714.88$, $p<0.0001$), as expected. Figure 4.4 illustrates a noticeable sub-trend (or potential sub-population) occurring below what was expected by the linear trend. This sub-trend suggests that higher populated ZIP codes expressed overall lower breast cancer risk. Figure 4.5 presents the residual output from the regression applied in Figure 4.4. Over- prediction of counts in larger populated zones is apparent in the residual pattern. Figure 4.6 provides a powerful cartographic visual of this phenomenon, illustrating US age-adjusted, ancestry-scaled breast cancer incidence during the study period. A noticeably lower breast cancer risk prevails throughout the majority of the Chicago region.

In Figure 4.7, ZIP code populations are regressed against ZIP code breast cancer counts with a best fit polynomial line expressing a slight quadratic trend. The negative quadratic trend suggests a parabolic effect of decreasing breast cancer counts as populations become large. Furthermore, the coefficient of determination for the quadratic best fit trend (Figure 4.7) was slightly improved from the linear best fit trend (Figure 4.4), as could have been anticipated.

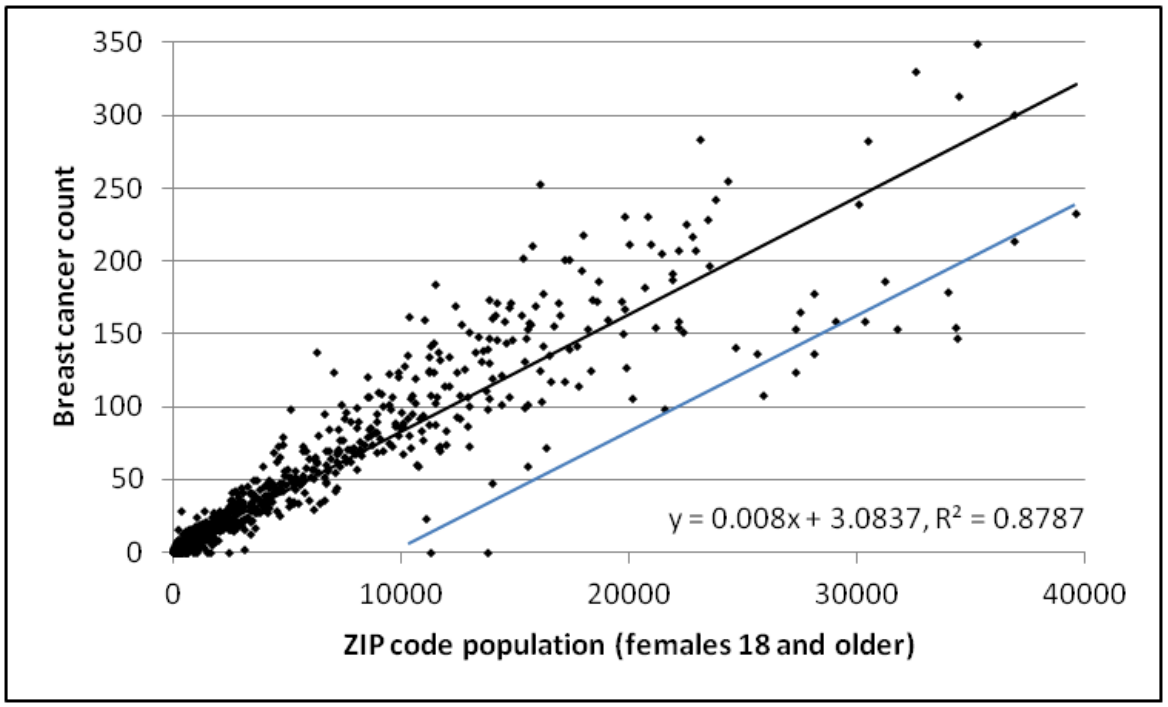


Figure 4.4. Bivariate regression of ZIP code population and ZIP code breast cancer counts. The blue line below illustrates a potential subpopulation.

Figure 4.5. Residual plot of ZIP code population regressed against breast cancer counts. The dotted line illustrates a trend of over-prediction occurring in higher populated zones.

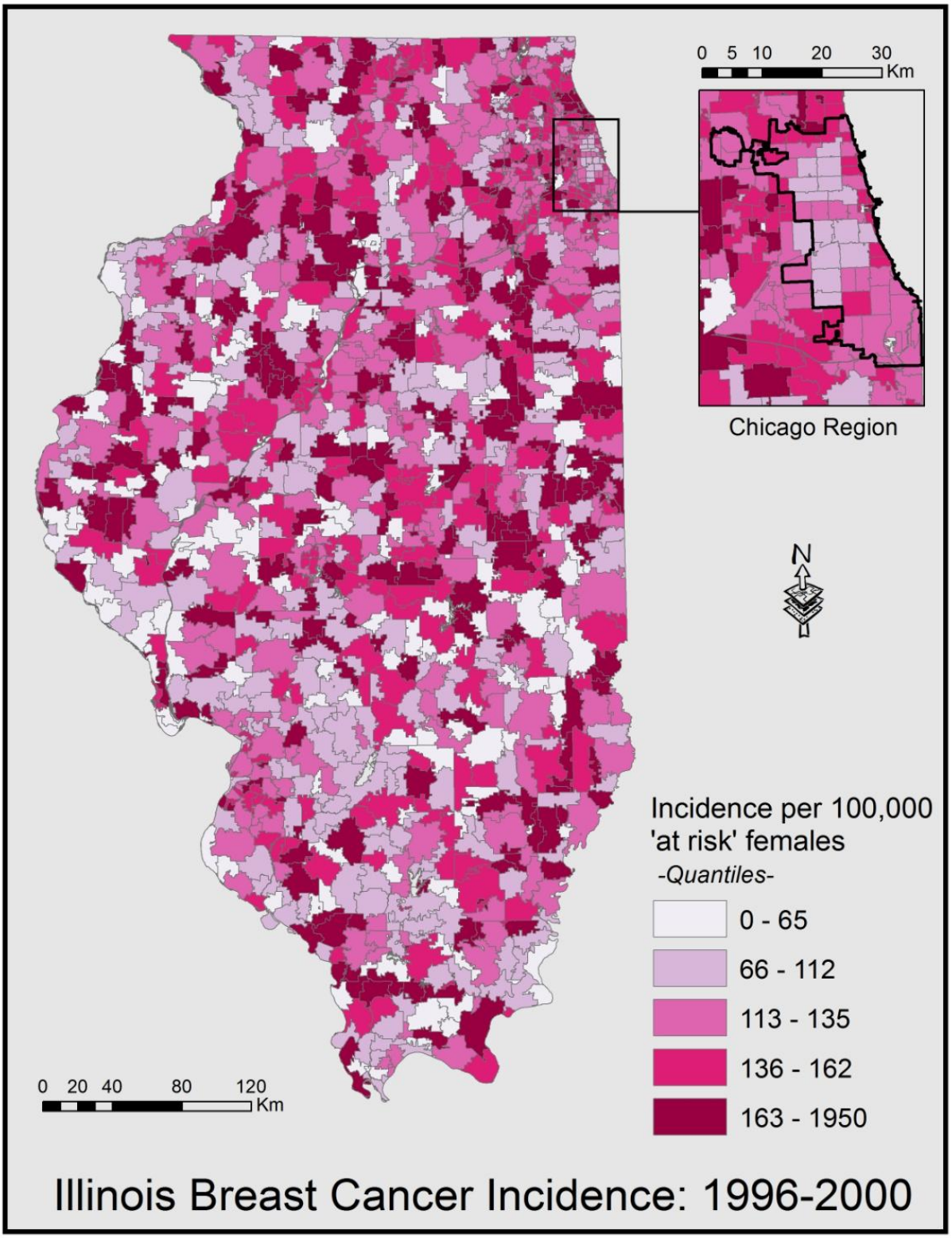


Figure 4.6. US age-adjusted, ancestry scaled breast cancer incidence (ages 18 and older) during the study period.

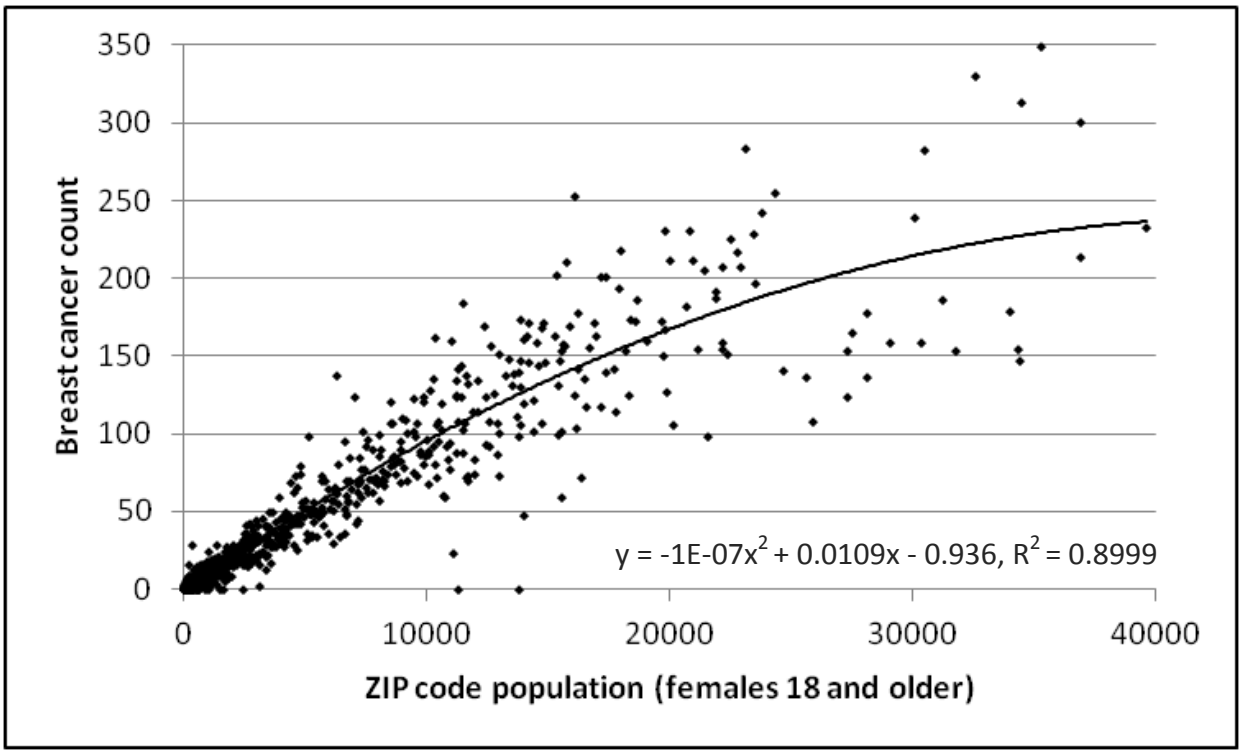


Figure 4.7. Bivariate regression of ZIP code population and ZIP code breast cancer counts, with a quadratic best fit polynomial line.

According to Figure 4.8, as the size of the female population increased, the ratio of Water Wells per female tended to asymptotically decrease, suggesting that Water Well exposure was predominantly a rural phenomenon. This is similar to the previously illustrated pattern of elevated breast cancer counts in rural ZIP codes. Perhaps there is an association between increased Water Well exposure and increased breast cancer incidence in rural zones.

The scatterplot of 'at risk' female population size and breast cancer incidence (Figure 4.9) expresses as strong message in terms of where risk gradients are located. According to this figure the majority of excess risk is located in lower populated zones.

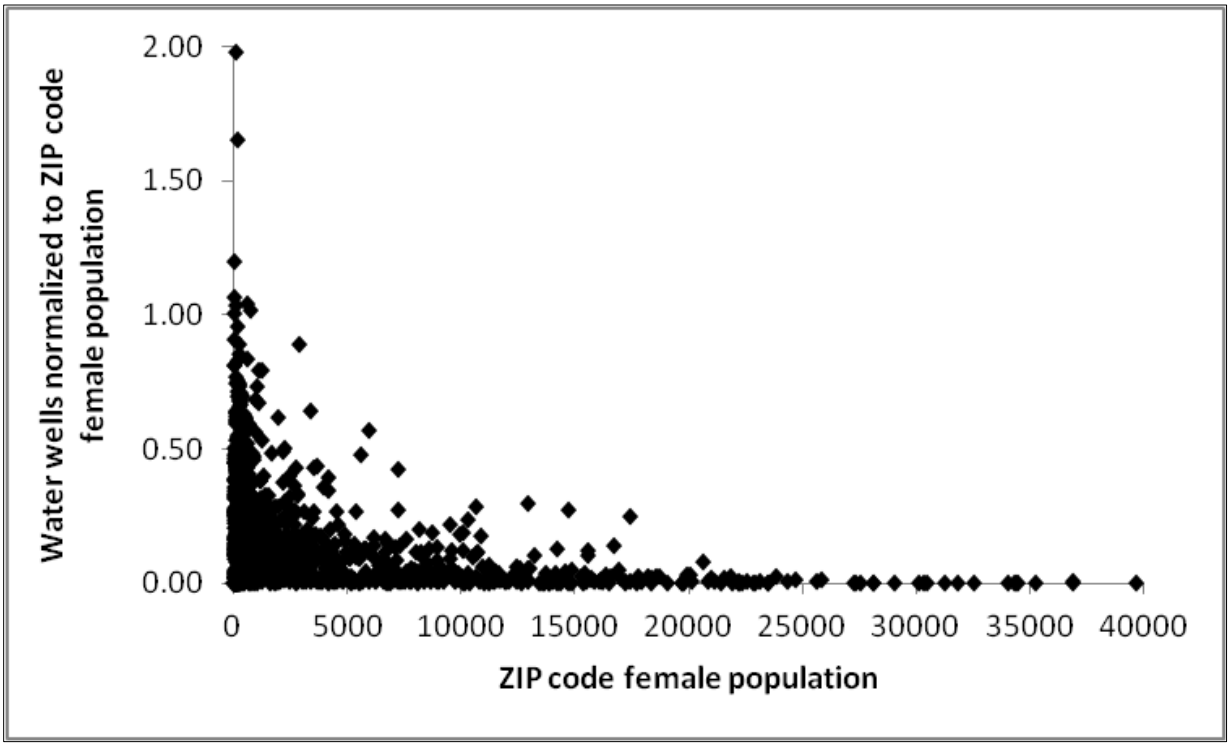


Figure 4.8. Ratios of Water Wells per female by ZIP code female population size.

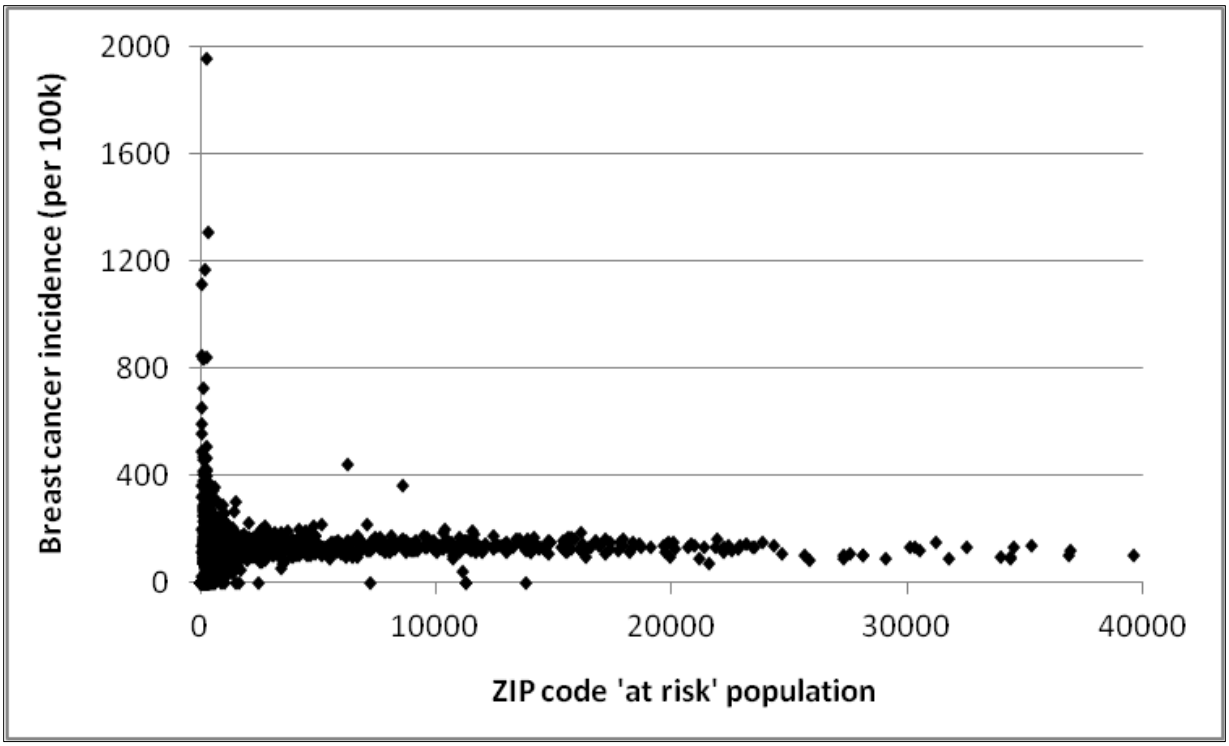


Figure 4.9. Breast cancer incidence by ZIP code ‘at risk’ population.

4.3 Clustering Test and Case/Control Design

Figure 4.10 shows the results from the Getis Ord Gi* cluster analysis. Results from this analysis culminated in the rejection of the null assumption of Poisson spatial variance. In total, 57 case ZIP codes and 674 reference ZIP codes were identified.

Table 4.2 illustrates EPA registered organizations that operated within Case ZIP codes during the cancer latency period. Some of these companies date back to the early 1900s, while others began operation in the 1980s. Since it takes many years for cancer to become clinically detectable, it is plausible to question the etiologic roles that these companies could have played in disease outcomes.

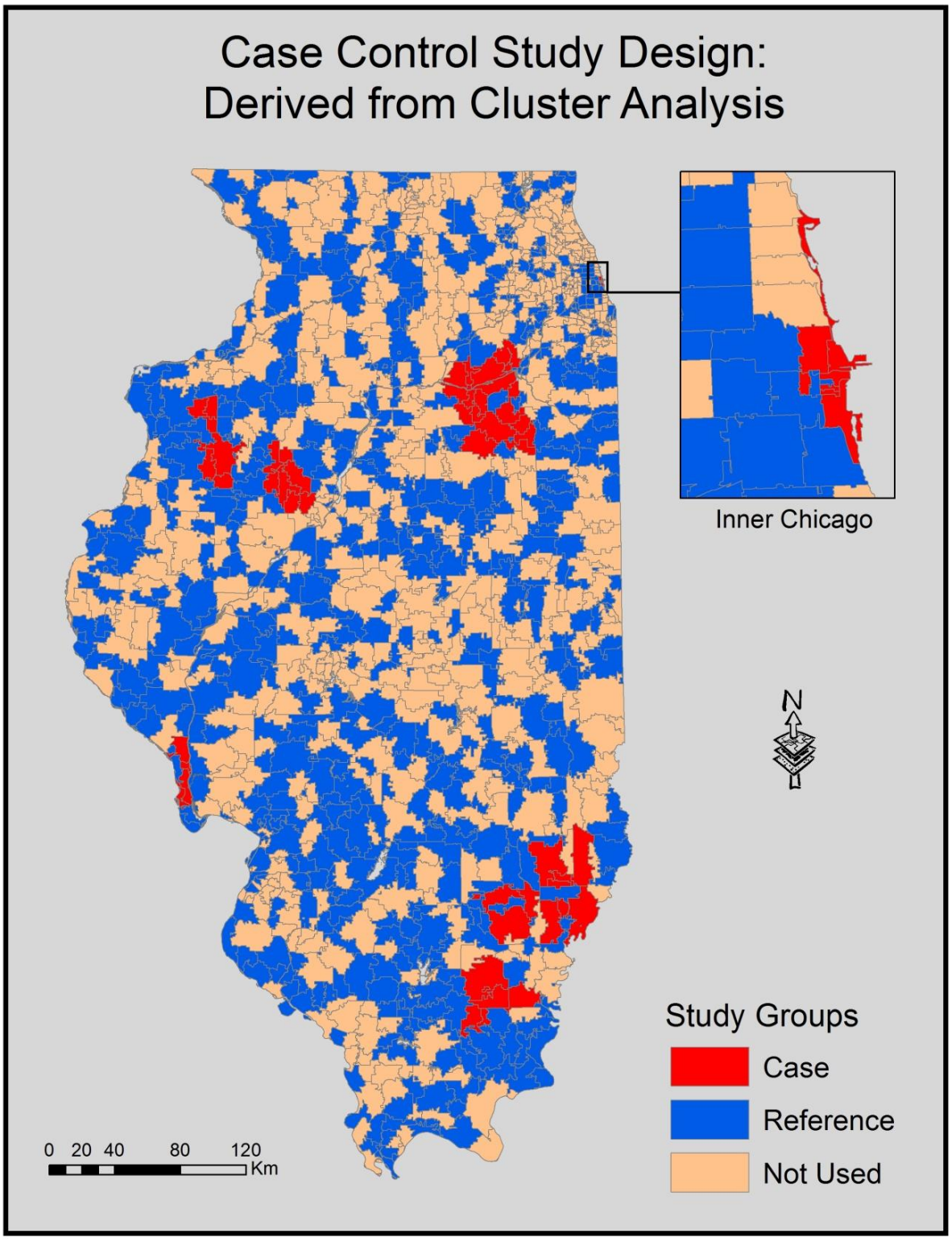


Figure 4.10. Case/control study surface derived from Getis-Ord Gi* cluster analysis. Map depicts 57 case ZIP codes and 674 reference ZIP codes.

Table 4.2. EPA registered facilities with environmental contamination potential. These facilities were located in Case ZIP codes and operated prior to the start of the study period.

ZIP code	City	Facility Name	X	Y	Contaminant*
60611	Chicago	Chicago Sun-Times	-87.6267	41.8892	3
60611	Chicago	Northwestern Memorial Hospital	-87.6215	41.8950	2, 3
60611	Chicago	Northwestern University, Chicago	-87.6201	41.8968	2
60611	Chicago	Rehabilitation Institute of Chicago	-87.6195	41.8958	2
60654	Chicago	Blommer Chocolate Company	-87.6428	41.8891	4
60654	Chicago	Watersaver Faucet Company	-87.6460	41.8934	4
61401	Galesburg	Archer Daniels Midland Company	-90.3837	40.9352	3, 4
61401	Galesburg	Butler Manufacturing Company	-90.3837	40.9352	3, 4
61401	Galesburg	Crop Production Services	-90.3977	40.9192	1
61401	Galesburg	Galesburg Sewage Treatment Plant	-90.4222	40.9401	5
61401	Galesburg	Inness Farm Supply, Inc.	-90.3447	40.9181	1
61401	Galesburg	Koppers, Inc.	-90.3958	40.8958	2, 4
61401	Galesburg	National Coatings Inc.	-90.3251	40.9264	4
61401	Galesburg	Sun Opta Ingredients	-90.3822	40.9306	4
61401	Galesburg	Tri States Water	-90.3834	40.9517	1
61402	Galesburg	Gates Corporation	-90.3414	40.9473	4
61402	Galesburg	Maytag Refrigeration Products	-90.3990	40.9351	3, 4
61341	Marseilles	Exelon-LaSalle Nuclear Power	-88.6645	41.2383	5
61341	Marseilles	Field Container Company LP	-88.7096	41.3263	4
61341	Marseilles	Marseilles Sewage Treatment Plant	-88.7222	41.3296	5
60447	Minooka	Grainco FS, Inc.	-88.2617	41.4600	1
60447	Minooka	Minooka Sewage Treatment Plant	-88.2389	41.4398	5
60450	Morris	Akzo Nobel Chemicals, Inc.	-88.3198	41.4175	1
60450	Morris	Akzo Nobel Surface Chemistry LLC	-88.3360	41.4065	2, 3, 4
60450	Morris	Exelon-Dresden Nuclear Power	-88.2686	41.3807	2, 5
60450	Morris	Midwest Generation LLC	-88.3404	41.3447	3, 4
60450	Morris	Morris Community Landfill	-88.4026	41.3714	3
60450	Morris	Northfield Block Company	-88.3643	41.3879	2
60450	Morris	Technical Propellants, Inc.	-88.2963	41.3857	4

*1=Pesticide Producer, 2=Hazardous Waste Generator, 3=Air Release Facility, 4=Toxic Release Inventory, 5=Wastewater Discharge

4.4 Statistical Analysis

Generalized Linear Mixed Model

Proc GLIMMIX in SAS 9.2 was used to evaluate the statistical association between case/reference ZIP codes and independent variables. In order to account for the spatial covariance within the model, three separate spatial covariance structures (Spherical, Exponential, and Gaussian) were evaluated using model fitness statistics. According to -2 Log Likelihood, Akaike Information Criterion, and Bayesian Information Criterion, the optimal spatial covariance structure was exponential decay (Table 4.2). The exponential spatial covariance formula is as follows (SAS 2012b):

$$\sigma^2 \exp(-d_{i,j} / \theta),$$

where negative distance (-d) is the ultimate governing factor in determining the intensity of covariance. As distance increases, the meaningfulness or size of the residual vector (θ) is reduced exponentially. At smaller distances, the residual vector is more influential.

Figure 4.11 illustrates the spatial covariance produced from analyzing the Case/Control design with ordinary Kriging (fitted very well with exponential decay). In further support of an exponential fitting, Ezra *et al.* (2006) and Jakhani *et al*. (2009) suggest that contaminant attenuation tends to be first-order linear up to 300 to 1000 meters from contaminant point sources. Beyond these distances, attenuation accelerates dramatically with further increases in empirical distance. Given the nature of exponential pollution decay, model fitness statistics (Table 4.3), and modeled spatial covariance (Figure 4.11), it was considered wise to implement an exponential spatial covariance.

The robust GLMM output (Table 4.4) expresses positive risk associations between breast cancer risk and Power Plants, Corn & Soybean Land Cover, Water Wells, Oil/Gas Wells, Hazardous Waste Generation Sites, and Mineral Borings. The cartographic results of the robust GLMM (Figure 4.12) illustrate the probability of breast cancer risk in association with environmental risk factors. Among these positive risk associations, Oil/Gas Wells, Mineral Borings and Large Quantity Hazardous Waste Generators express high statistical significance in association with increased breast cancer risk. Pesticide Producing Facilities, Wastewater Releases, Oil/Gas Injection Wells, and Oil/Gas Storage Observation Sites, express a negative association with breast cancer risk. Among these negative associations, Pesticide Producing Facilities expressed high statistical significance in association with reduced breast cancer risk.

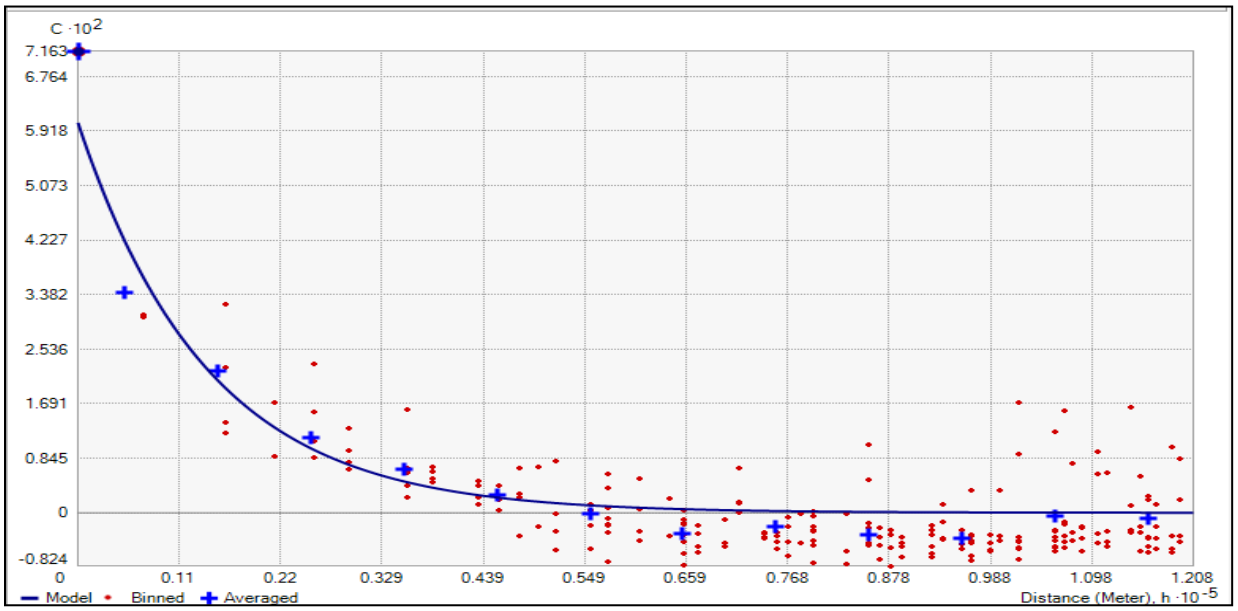


Figure 4.11. Exponential line fitted through the case/control model's spatial covariance (using ordinary Kriging on the binary response variable).

The model fitness statistics for the reduced GLMM (Table 4.5) suggest a similar pattern within spatial covariance. Results from the reduced GLMM (Table 4.6) display nearly identical results in comparison to the robust model, with regard to the significant variables. Hazardous Waste Generation Sites, Oil/Gas Welling Sites and Mineral Borings each displayed high statistical significance in association with increased breast cancer risk. Whereas, Pesticide Producing Facilities displayed high statistical significance in association with reduced breast cancer risk. The interactions between significant covariates from the reduced model were all non-significant (Table 4.7).

Table 4.3. Robust GLMM fitness statistics.

Fitness Statistics			
Spatial Covariance Structure	-2LogL	AIC	BIC
Spherical	158.7	160.7	165.25
Exponential*	65.4	69.4	78.55
Gaussian	141.2	145.2	154.4
None	158.7	170.7	198.1

**Best covariance structure*

Table 4.4. Model solution from the robust GLMM.

GLIMMIX Model Solution †				
Effect	Estimate	Std. Error	t Value	p Value, <
Intercept	0.015890	0.022050	0.72	0.4715
Power Plants	0.122900	0.071570	1.72	0.0863
Water Wells	0.070500	0.053230	1.32	0.1857
Corn/Soybean Land Cover	0.049540	0.037290	1.33	0.1844
Oil/Gas Wells	0.000041	0.000009	4.40	0.0001**
Hazardous Wastes	0.000029	0.000008	3.59	0.0003**
Mineral Borings	0.000029	0.000006	4.93	0.0001**
Oil/Gas Injection Wells	-0.000040	0.000303	-0.12	0.9015
Pesticide Producing Facilities	-0.000050	0.000019	-2.46	0.0142*
Wastewater Releases	-0.000100	0.000073	-1.32	0.1881
Oil/Gas Storage Observation Sites	-0.000210	0.000347	-0.61	0.5400

*†Exponential spatial covariance structure, **Highly Significant, *Significant*

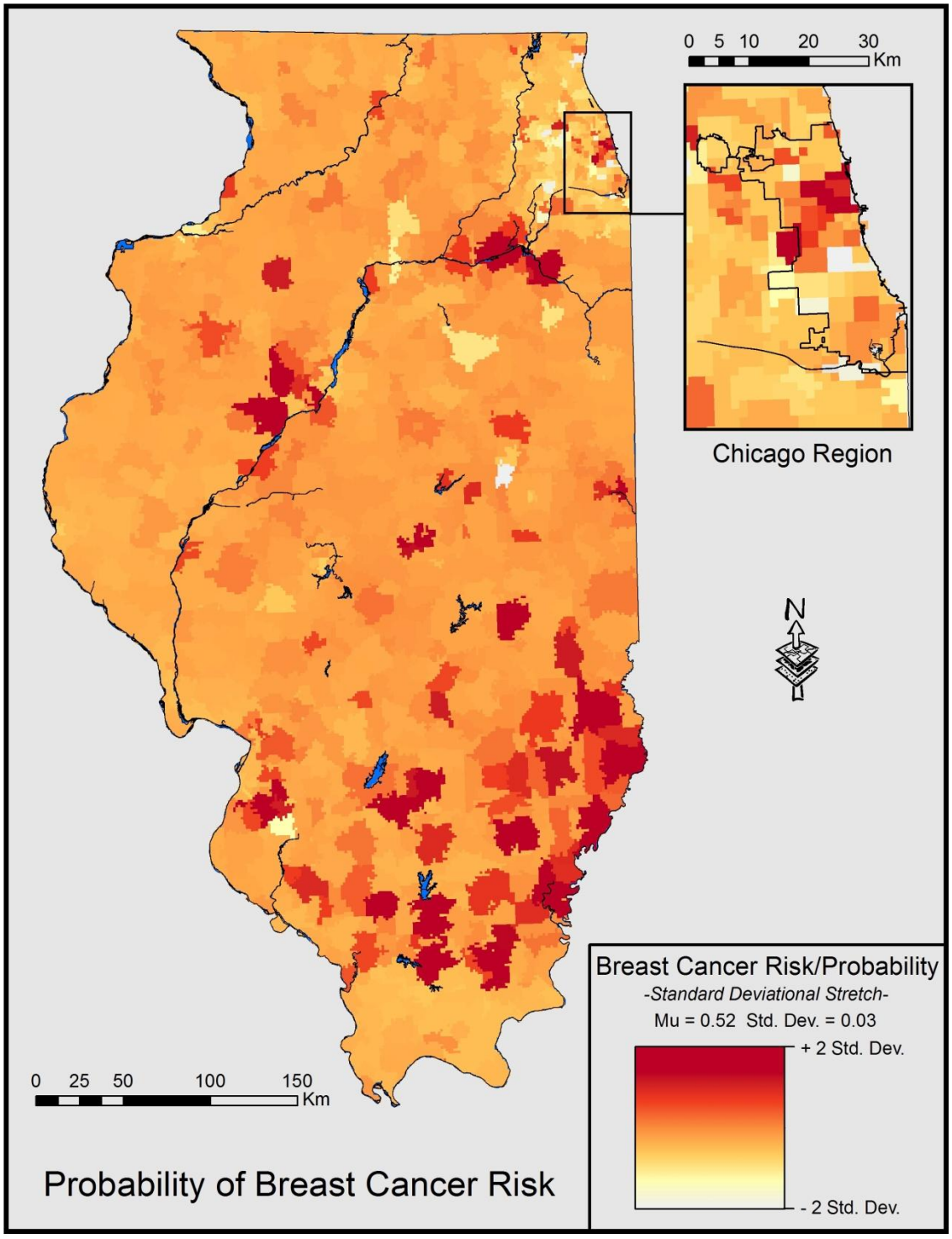


Figure 4.12. Breast cancer risk prediction map resulting from the robust GLMM. Model processed through ArcMap's Raster Calculator.

Table 4.5. Reduced GLMM fitness statistics.

Fitness Statistics			
	Fit Statistics		
Spatial Covariance Structure	-2LogL	AIC	BIC
Spherical	65.4	77.4	104.9
Exponential*	-28.1	-14.1	17.9
Gaussian	47.8	61.8	93.9
None	65.4	77.4	104.9

**Best covariance structure*

Table 4.6. Model solution from the reduced GLMM.

GLIMMIX Model Solution †				
Effect	Estimate	Std. Error	t Value	p Value, <
Intercept	0.056060	0.010130	5.53	0.0001**
Oil/Gas Wells	0.000040	0.000009	4.41	0.0001**
Mineral Borings	0.000029	0.000006	4.97	0.0001**
Hazardous Wastes	0.000024	0.000008	3.10	0.0020**
Pesticide Producing Facilities	-0.000050	0.000019	-2.62	0.0089**

*†Exponential spatial covariance structure, **Highly Significant*

Table 4.7. Interactions within the reduced model.

Type 3 Tests of Fixed Effects		
Effect	F value	p value, <
Hazardous Wastes * Mineral Borings	3.29	0.0701
Hazardous Wastes * Oil/Gas Wells	0.35	0.5571
Hazardous Wastes * Pesticide Producers	0.62	0.4306
Mineral Borings * Pesticide Producers	1.42	0.2346
Mineral Borings * Oil/Gas Wells	1.75	0.1866
Oil/Gas Wells * Pesticide Producers	1.26	0.2611

Discriminant Function Analysis

The multivariate F-tests of differences between group means proved to be highly significant (Table 4.8), suggesting that prediction model has the ability to distinguish between risk zones. Furthermore, the differences of means across groups were statistically significant for all variables (Table 4.9), with high statistical significance occurring in Black Population, Asian Population, High School Diploma and Managerial Employment. Hispanic population was slightly significant at the 0.05 α level. The spatial distributions for race by risk zone can be observed in Appendix B. Race was effective at discriminating between the Chicago risk zone and the remainder of risk zones, suggesting that race is an important factor when comparing urban disease outcomes incidence to rural disease outcomes, in Illinois.

From analysis of group means and standardized function coefficients (Table 4.10), the five predictor variables functioned as a contrast. Black Population, Asian Population, and Managerial Employment worked strongly together as group distinguishing factors. Hispanic Population and Without High School Diploma worked mildly together as group distinguishing factors. The contrast between these two sets of variables suggests that there could be two socioeconomic linkages within the model. One of the linkages appears to be between race and employment, and the other appears to be between race and education. The linkage between race and employment expresses the strongest canonical association.

According to Table 4.11, the discriminant function was good at classifying ZIP codes into risk zones 3 and 5. It performed mediocre with classification back into risk zones 2 and 4, and the reference group. The discriminant function performed poorly with regard to classification back into risk zone 1.

Cartographically, the association between Managerial Employment and cancer risk is most apparent in the northern zones of Illinois, with the Chicago zone expressing the strongest managerial gradient (Figure 4.13). Similarly, higher academic achievement expresses a strong association with breast cancer risk in the northern portion of the state (Figure 4.14), while breast cancer risk in southern Illinois is associated with lower academic achievement.

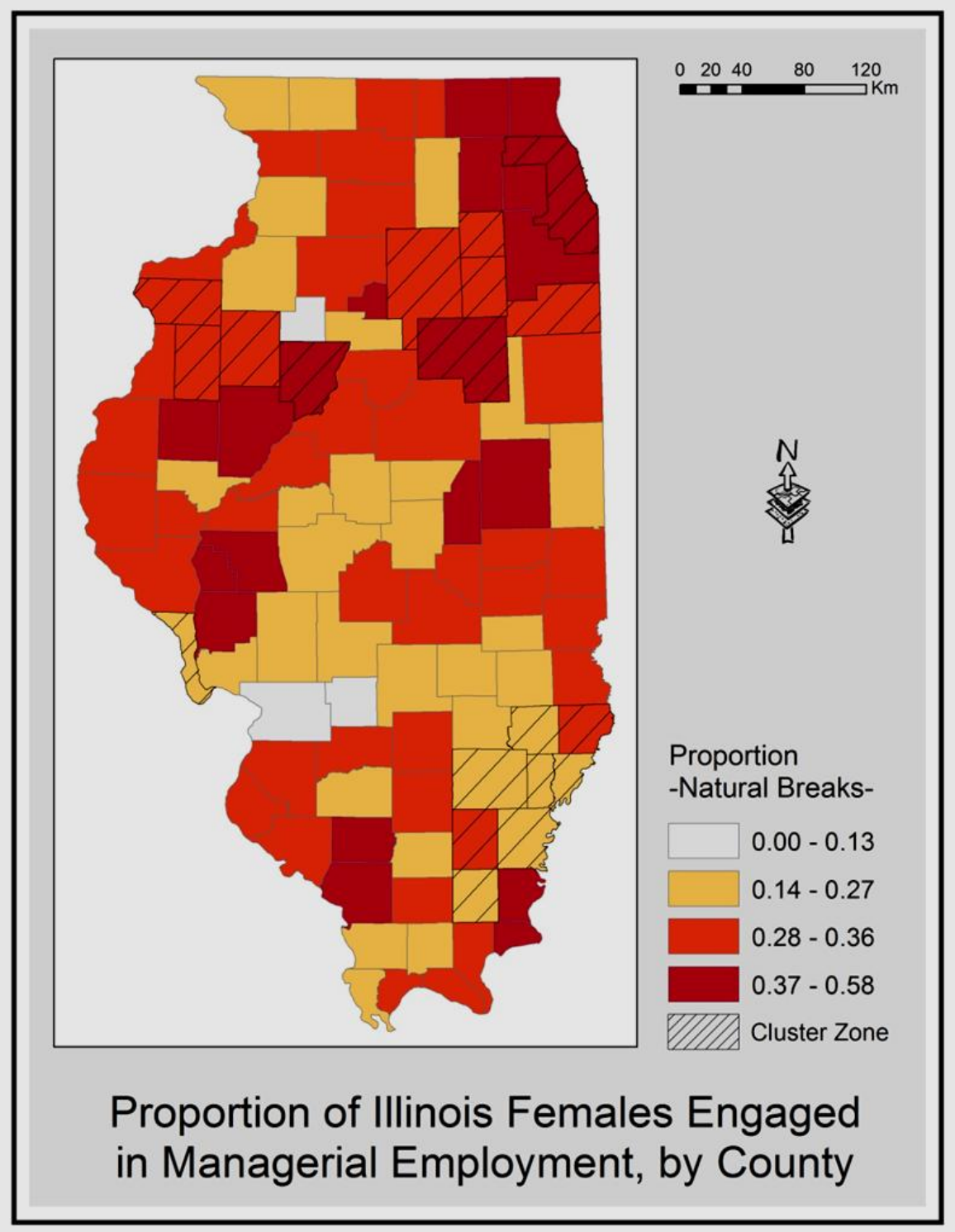


Figure 4.13. The proportion of Illinois females engaged in managerial employment. The map also illustrates modeled breast cancer risk.

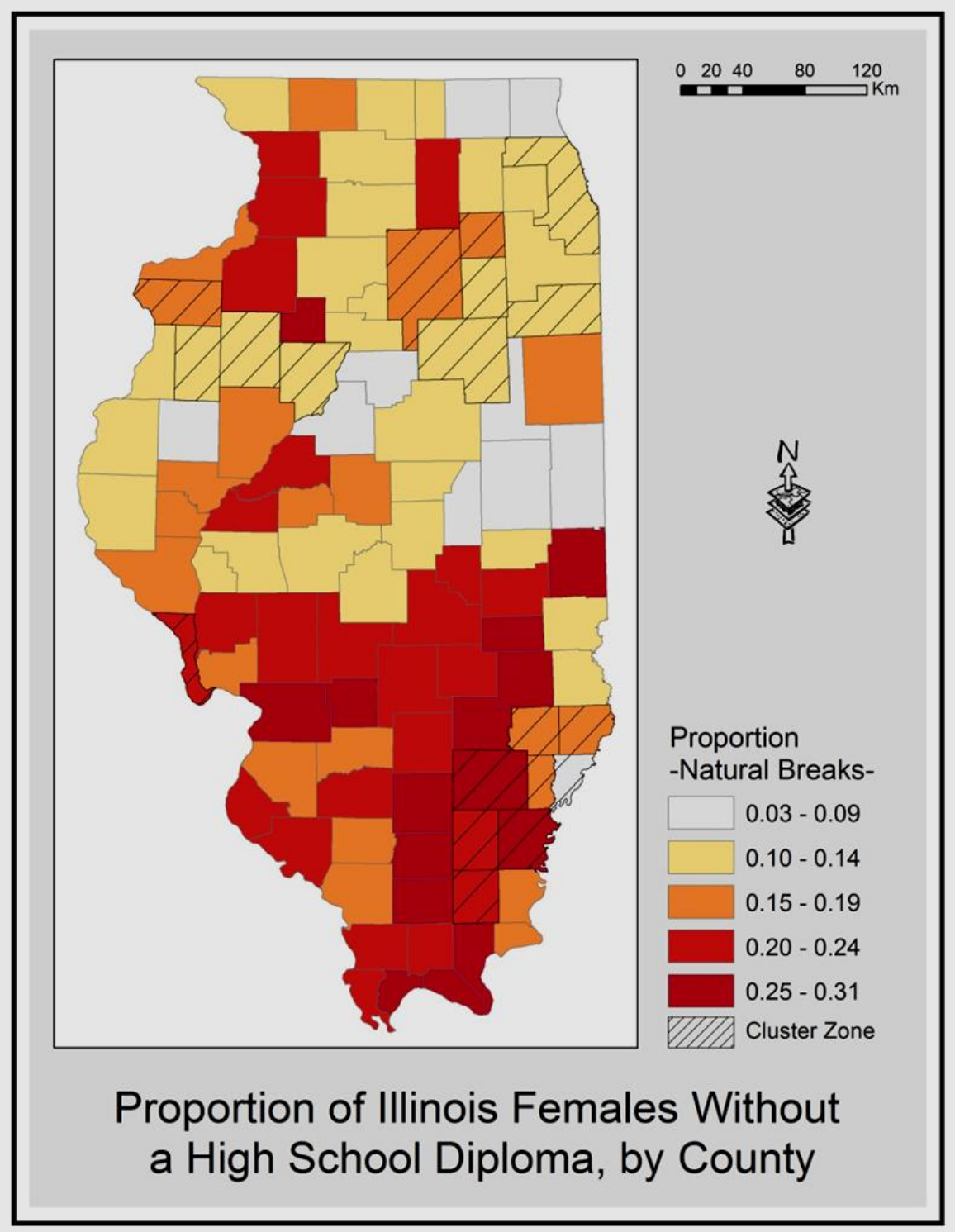


Figure 4.14. The proportion of Illinois females who do not possess a high school diploma. The map also illustrates modeled breast cancer risk.

Table 4.8. Multivariate tests of mean differences from the Discriminant Function Analysis.

Multivariate Test of Mean Differences			
Statistic	Value	F-value	p-value, <
Wilks' Lambda	0.5258	5.71	0.0001**
Pillai's Trace	0.5387	5.00	0.0001**
Hotelling-Lawley Trace	0.7827	6.32	0.0001**
Roy's Greatest Root	0.6032	24.97	0.0001**

***Highly significant*

Table 4.9. Univariate test for variable mean differences across groups.

Univariate Test Statistics			
Predictor Variable	R-Square	F-value	p-value, <
Black Population (%)	0.1886	9.62	0.0001**
Hispanic Population (%)	0.0574	2.52	0.0306*
Asian Population (%)	0.3000	17.74	0.0001**
Without High School Diploma (%)	0.1287	6.12	0.0001**
Managerial Employment (%)	0.1324	6.32	0.0001**

*** Highly significant, *Significant*

Table 4.10. Analysis of group means and standardized discriminant coefficients.

Group Means & Standardized Function Coefficients							
	Reference	Risk Zone Group Mean (%)					Function
Predictor Variable	Group Mean (%)	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5	Coefficient*
Black Population	4.69	1.32	0.11	16.45	0.24	0.00	0.5673
Hispanic Population	3.90	1.11	2.86	5.20	0.63	1.18	-0.1497
Asian Population	0.81	0.24	0.17	11.01	0.34	0.10	0.5709
Without High-School Diploma	18.62	11.42	12.33	4.33	19.96	23.78	-0.4934
Managerial Employment	30.04	30.73	28.27	54.71	31.26	28.20	0.6220

**Derived from a highly significant canonical correlation (F=5.74, p<0.0001)*

Table 4.11. Classification summary from the Discriminant Function Analysis.

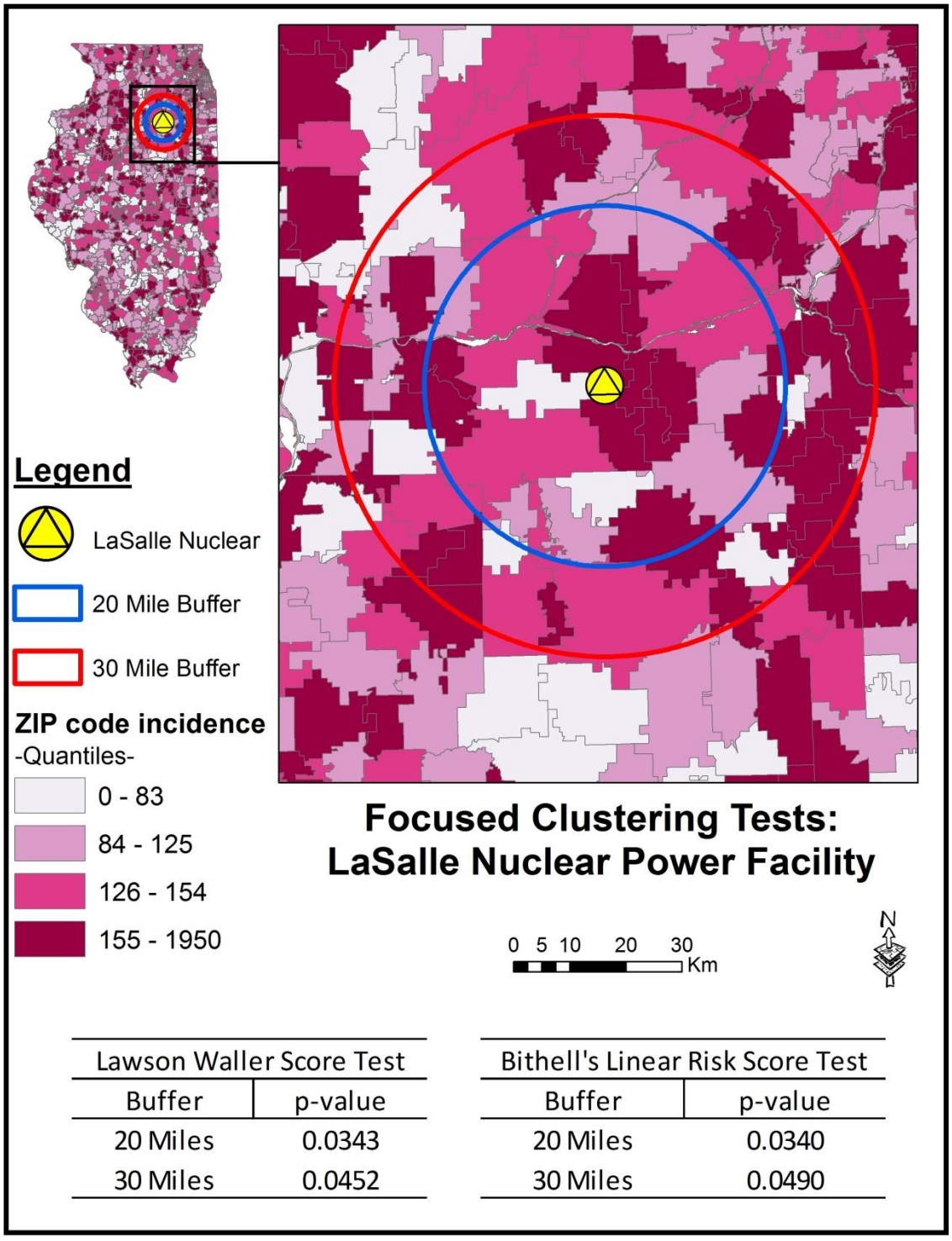
Resubstitution Summary								
	Classification							
From group	Reference	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5	Total	% Correct
Reference	102	5	14	0	31	9	161	0.63
Zone 1	6	3	0	0	5	0	14	0.21
Zone 2	2	0	8	0	1	2	13	0.62
Zone 3	0	0	0	6	0	0	6	1.00
Zone 4	4	0	1	0	8	2	15	0.53
Zone 5	0	0	0	0	0	5	5	1.00
TOTAL	114	8	23	6	45	18	214	0.62
PERCENT	0.53	0.04	0.11	0.03	0.21	0.08		
PRIORS	0.75	0.06	0.06	0.02	0.07	0.02		
RATIO	0.71	0.62	1.79	1.40	3.00	4.21		

4.5 Spatial Analysis

Focused Clustering Tests

Focused clustering tests were conducted in various areas expressing gradient levels of contaminant exposure. Results from focused tests suggest elevated breast cancer risk near the LaSalle Nuclear Power Facility (Figure 4.15) and near ZIP codes 61529, 61536 and 61569 (Figure 4.16).

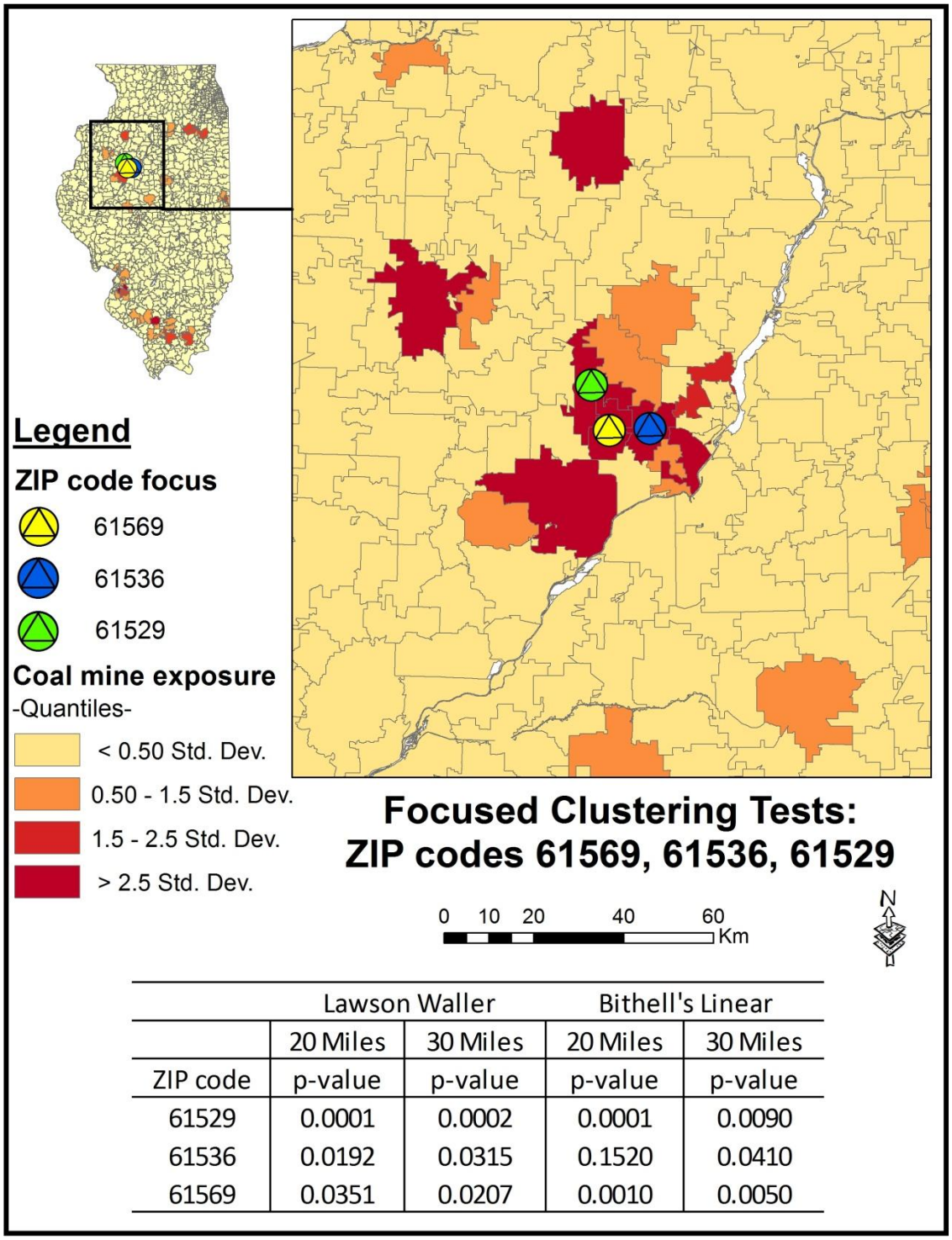
This LaSalle Nuclear Power Plant became operational in 1982, fourteen years prior to the opening of the study window. ZIP codes 61529, 61536 and 61569 were selected for focused clustering tests due to extreme frequencies of mineral borings.



Lawson Waller Score Test	
Buffer	p-value
20 Miles	0.0343
30 Miles	0.0452

Bithell's Linear Risk Score Test	
Buffer	p-value
20 Miles	0.0340
30 Miles	0.0490

Figure 4.15. Focused clustering tests surrounding the LaSalle Nuclear Power Facility.



	Lawson Waller		Bithell's Linear	
	20 Miles	30 Miles	20 Miles	30 Miles
ZIP code	p-value	p-value	p-value	p-value
61529	0.0001	0.0002	0.0001	0.0090
61536	0.0192	0.0315	0.1520	0.0410
61569	0.0351	0.0207	0.0010	0.0050

Figure 4.16. Focused clustering tests surrounding the locations of ZIP codes 61569, 61536 and 61529. This area represents the most intense zone of coal mining activity for the entire state of Illinois.

CHAPTER 5

DISCUSSION

Environmental Risk Patterns

Breast cancer clustering occurred in four distinct geographic zones. These zones consisted of the Lower Illinois River Basin (LIRB), Upper Illinois River Basin (UIRB), inner Chicago, and in the southeastern area of the Illinois Basin.

The LIRB and UIRB are host to environmental risk factors and carcinogens as highlighted in Brody and Rudel (2003), Davis *et al.* (1998), Gray (2010), Groschen *et al.* (2000), Huggins *et al*. (2009), ISGS (2010c), Jüngten and Klein (1977), Kolyoncu (1997), Korte and Fernando (1991), Morrow (1999), NAWQA (1994), NIH (2002, 2011), Rudel *et al.* (2007), Voldner and Li (1995), Warner (2001), Warner *et al.* (2003),Welch *et al.* (1988) and Wolff and Toniolo (1995). These risk factors include volatile organic compounds (e.g. chloroform, MTBE, BTEX), dioxins, organochlorines and soil fumigants (e.g. PCBs, DDT, nitromethane, dieldren), herbicides (e.g. atrazine, alachlor), commercial hazardous wastes, inorganic compounds (e.g. arsenic and radium contaminated aquifers) and toxic metals from mining (e.g. mercury, manganese, etc).

In the LIRB, just to the west of Peoria, IL (ZIP codes 61517, 61569, and 61529) contained 373, 2434, and 4411 historic Mineral Borings, accordingly. The 75th percentile for Mineral Borings was seven (7) throughout the entire state. This research illustrated that despairingly high frequencies of Mineral Borings associate with increased breast cancer risk. Slightly further to the west (in Galesburg, IL), in ZIP codes 61401 and 61402 displayed a wide

variety of commercial environmental risks. Toxic Release Inventories and Air Release Facilities accounted for the majority of exposure risk within the Galesburg area.

Within the UIRB, approximately 75 to 100 kilometers southwest of Chicago, an extreme zone of risk emerged near the confluence of the Des Plaines and Kankakee Rivers. This area included ZIP codes mostly within LaSalle, Grundy, and Livingston Counties. The dominant commercial cities in this risk area appeared to be Morris, Minooka, and Marseilles, IL. This zone is host to Nuclear Power Plants, Corn & Soybean Land Cover, Mineral Borings, Hazardous Waste Generators, Wastewater Discharges, and Pesticide Producers. Three (3) of the six (6) Nuclear Power Plants in Illinois were located in this risk region; furthermore, the LaSalle Nuclear Facility tested significantly for increased focused risk at both a 20 and 30 mile bandwidth. In many ways, this region expressed a unique environmental risk portfolio, where most types of environmental risk were present at high levels.

The southeastern portion of the Illinois Basin expressed high risk associated with Oil/Gas Welling, Injection Welling, Corn/Soybean Agriculture, and Coal Mining. The ZIP codes involved in this association were members of Saline County, Hamilton County, White County, Wayne County, Edwards County, Wabash County, Richland County, and Lawrence County. Relevant to this zone of risk, Rudel *et al.* (2007), Gray (2010), ATSDR (1999), EPA (2010d), ISGS (2010c), Jüngten and Klein (1977), and Huggins *et al.* (2009), outline a variety of associated chemical risk factors such as polycyclic aromatic hydrocarbons, heavy metals, radioactive isotopes, and inorganic and organic compounds, that compliment risk factors identified by the GLMM. Overall, the GLMM indicated that this region was highly susceptible to risk associated with hydrocarbon exposure, due to the high prevalence of Oil/Gas Welling. EPA literature (EPA 2010d) suggests that chronic exposure to radon is associated with oil

welling in the Midwestern United States. It is possible that hydrocarbon exposure in the midst of radioactive isotope and heavy metal exposure increases the risk that occurs in this zone.

In the Chicago zone of risk, there was a high likelihood of exposure to Hazardous Waste Generators, Air Release Facilities, Toxic Release Inventories, as well as nearby Power Plants. The per capita potential for exposure was high in Chicago ZIP codes, given the drastically elevated population densities. According to the GLMM, Hazardous Wastes provided the most risk association in Chicago. One specific chemical, Ethylene Oxide, a noted breast cancer risk factor (Rudel *et al.* 2007), appeared to be a chemical risk in the Chicago zone via medical equipment exposure and intense vehicle exhaust exposure. Additional occupational risk in this zone could be related to chronic radioactive (x-ray) exposure at the previously mentioned health treatment facilities.

Hydrologic Trends

Visual analysis of the Case/Control Study Area displayed that breast cancer clusters occurred in close proximity to major water sources. These water sources included the Illinois River, the Des Plaines River, the Kankakee River, the Mississippi River, the Wabash River, and Lake Michigan. Three of the five risk zones occurred near river confluences (the Des Plaines and Kankakee Rivers, the Illinois and Mississippi Rivers, and the Wabash and Ohio Rivers). This suggests a potential systematic pattern of risk where rivers merge. It has already been noted that the Illinois River functions as a transportation network for Chicago's wastes and refuse (NAWQA 1994). Pollution of the Illinois River could intensify as tributary waters leave the Chicago area and eventually merge into the Illinois River –and ultimately continue toward the Mississippi River. The same principal could apply as the Wabash River flows toward the

confluence with the Ohio River. Overall, it appeared that the greatest proximal risk was associated with the Illinois River system and the Wabash River. Similar hydrologic associations were observed by Guajardo and Oyana (2009) during an assessment of breast cancer clusters near the Tittabawassee and Saginaw Rivers, in Michigan.

Socioeconomic Trends

The vast majority of breast cancer risk occurred in rural zones accompanied mostly with lower economic status and lower educational status. Economic status could be a non-contributing factor, given that the costs of living could be adjusted to account for differences between urban and rural zones. Education appears to be prevalently low across the southern half of the state Illinois, with increasing education attainment closer to Chicago. New research questions could be hypothesized from this observation. For instance, is education associated with personal behaviors such as smoking, alcohol consumption and obesity (breast cancer risk factors)? Common rhetoric about breast cancer risk states that higher socioeconomic status (SES) commonly accompanies greater risk (Brody *et al.* 2003). Breast cancer risk within the Chicago zone supported this opinion, because higher education and higher employment status were expressed in the Chicago zone. Racially, the risk in Chicago appeared to be diverse.

This research does not fully support the argument that higher socioeconomic status correlates with increased breast cancer risk. Instead, this research suggests that breast cancer risk is a problem across a variety of socioeconomic statuses, some higher or lower than others. Breast cancer risk expressed a strong association with Black and Asian Populations within urban communities; whereas, in rural communities, breast cancer risk associated with lower economic/employment status and low educational attainment.

The highest human exposure to hazardous waste generation does occur in the Chicago area. If hazardous waste exposure was more evenly dispersed across Illinois, then this exposure would not be unique to Chicago residents. However, it is unique to Chicago residents, which further supports the concept that Asian, Black, Higher Education, and Managerial Employment are associated with more intense zonal exposure to Hazardous Waste Generation Facilities.

Similarly, the discriminant function analysis suggested that exposure to Agricultural Chemicals, Coal Mining, Oil/Gas Welling, Pesticide Producers, Water Wells, and Power Plants was predominantly a rural phenomenon experienced by members of the White Population (this was not modeled statistically, but it is intuitive) who tended to be less engaged in Managerial Employment and also less educated.

Potential Confounding Effects

It is possible that the higher incidence in rural Illinois is a result of 'the small population problem'. Different incidence filtering techniques (i.e. Ordinary Kriging, Bayesian Smoothing, Spline Interpolation, etc) are commonly used when such a problem is suspected; however, these techniques can induce excessive information borrowing (over-smoothing) and can complicate Poisson assumptions, if such assumptions are meaningful. The best alternative might be to extend the study temporally and look for consistencies within apparent 'small population' zones (Jacquez 2011). Temporal consistencies could be used to support the idea of a true risk in a small population zone and help to bypass 'denominator-based quotient problems'.

It is the opinion of the researcher that the data observed in this analysis were prepared in a way that controlled for population effects, age effects, and ancestry associations, while still preserving a meaningful spatial estimate of disease risk.

Potential 'edge effects' might have influenced the clusters identified in the Chicago zone and the Calhoun County zone. ZIP codes located adjacent to water sources (just as these two zones were) tend to have fewer neighbors, causing neighborhood index contributions to be provided by fewer neighbors. Under these conditions, outliers can have the ability to spuriously cause clusters, particularly when areal adjacency is used to conceptualize spatial relationships within cluster analysis.

GLMM Power Analysis

It can be argued that the reduced model possesses a level of power that is able to reliably estimate regression coefficients. The 'k+1' value (number of independent variables plus the intercept) in the reduced model is less than one-tenth of the size of the smallest binary group (the Case group was smallest with n=57). The 'k+1' value of the reduced model (k+1=5) is less than one-tenth of the Case group size. Specifically, $[(k+1)/n_{(\text{smallest group})} < 0.10]$ or 5/57=0.0877. This approach can be referred to as a 'rule of thumb' or, perhaps, as a 'golden rule of one-tenth' (Hosmer and Lemeshow 2000; Peduzzi *et al.* 1996) that can be applied to generalized linear models for binary outcomes.

Zero Counts in the Disease Data

Bimodality was an initial challenge in approaching this disease dataset, since 13.6 percent of ZIP codes presented zero counts of breast cancer. This complicated the ability to achieve a random type of distribution needed to fully approach the research under a null assumption of Poisson variation. Disease data for this research could have been inaccurately reported (i.e. under-reported) by local physicians, county health departments, state offices, or etcetera,

causing an over-prevalence of zero counts and a subsequently clumped distribution. In instances as these it can prove beneficial to employ a risk filter that estimates disease incidence based off of neighboring values (see Table 4.1 to observe the effects of employing an empirical Bayesian filter), thus reducing zero count prevalence and narrowing the gap between variance and mean.

CHAPTER 6

CONCLUSION AND RECOMMENDATIONS

Future research should attempt to include longitudinal cohort data or geographic lifeline information. Such data would partially account for individual dwelling times and help to validate the timing and length of exposure to environmental risk factors. The ZIP code scale of resolution within this study complicates attempts to provide more precise details about variables. It would be beneficial in future research to target the five zones that have been identified in this research and seek to synthesize additional confounding information. The task of synthesizing would be made easier if future research is afforded higher resolution disease data (i.e. census tract or individual level data).

In many ways, this research is a good example of how a variety of spatial data can be utilized within a scientific research design to draw spatial inferences. For instance, the Illinois wells and borings dataset contained 52 Point Feature Classes and over 500,000 records. It is unknown whether the ISGS, USDA-NASS, or EPA intended for their datasets to be included in a spatial epidemiologic study of breast cancer. The power of data mining, imagination, scientific understanding, literature review, and spatial and statistical analysis, have the ability to make potentially arbitrary data useful. The key factor is geocoding and georeferencing the data elements, in order to evaluate them spatially.

The focused clustering techniques stand to be criticized. The scale effects of ZIP code data prevent the employment of a biologically meaningful attenuation buffer in rural areas. For instance, the large sizes of rural ZIP codes require a radial buffer of 20 miles from a focused point of exposure, in order to achieve a sample size of $n > 30$. On the other hand, the nature

radioactive attenuation would compliment a 20 mile buffer. Therefore, focused tests around Nuclear Power Plants are sensible. However, exposure to hydrocarbons and inorganic/organic chemicals would likely be physiologically non-responsive past one or two miles of radial buffer.

There are still confounding factors and random effects that need to be explained in this study surface. This represents another downfall of large, aggregated data (i.e. ZIP code data). Individual level data would undoubtedly resolve the issue of 'unobserved confounding' effects.

With regard to using Proc GLIMMIX in SAS 9.2, it should be noted that raw UTM coordinates produced convergence problems due to infinite likelihoods. Thus, a UTM conversion factor was applied to crude coordinates. UTM (in meters) was multiplied by 0.0025, giving a geometric conversion close to kilometers (3/4 km) that allowed SAS to converge after four or five iterations.

Furthermore, future studies should evaluate how SAS can implement spatial referencing systems such as UTM 16N within a GLMM. The GLMM within this research was applied to a non-projected surface or, at best, to a "smooth spheroidal" world. It is unknown how the geodesic conceptualization of space impacted standard errors and regression coefficients. Perhaps these effects would be unnoticeable. Future research should consider geodesic conceptualizations within SAS.

Future Directions

The IDPH, ISGS, ISWS, IDNR, ILEPA, USEPA, and CDC, should be aware that environmental breast cancer risk (when population standardized) is unequivocally a rural problem in Illinois. The long history of agricultural industry, fossil fuel mining, private water welling, waste transportation, power generation, and chemical production and subsequent

discarding, is a story about rural geography. In addition to increased environmental risk, rural Illinois is also host to a higher proportion of older age females who are more susceptible to neoplasia

Future environmental risk assessments within Illinois should attempt to account for behavioral risk factors and other intrinsic risk factors. Behavioral factors should include tobacco smoking, alcohol consumption, exercise frequency and breast feeding. Intrinsic factors should include body fat composition (obesity), genetic predisposition, breast feeding, age of menarche/menopause and viral disease encounters during lifetime. It would be beneficial to observe chemical profile information within high risk zones, looking specifically at soil constituents, dissolved chemicals, water quality, and human tissues. A thorough investigation of clinical data would clarify many questions about disease Cases. Length of residency within risk zones should also be evaluated. This would help to determine the likelihood of exposure to modeled putative sources during the cancer latency period.

If individual level data is not obtainable in future research, then it would behoove of the researcher to employ spatial computing algorithms to develop disease surfaces that are more biologically meaningful than political enumeration districts such as ZIP codes, census tracts, etc. This is suggested since there is little to no inherent relationship between political boundaries and physiological outcomes.

The modifiable areal unit problem remains an issue with areal based spatial analysis, such as the events of this research. Political boundaries are typically non-stationary over time. Periodic changes in political zoning can associate with shifts in disease intensity, thus spuriously causing the relocation of disease clusters (Lawson 2006). Thus, it again behooves of future researchers to utilize computing algorithms that can calibrate best conceptual models of

spatial boundaries, based on combinations of physiologically meaningful covariates. These variables should come from biographical, clinical, biological, and chemical data. A subsequently related challenge would be to communicate research findings to state policymakers, because algorithmically synthesized disease areas might prove to be conceptually vague to politicians and non-scientists.

Lastly, the need for individual-level disease data and high resolution census data cannot be overstated. These types of higher resolution data elevate the certainty of research findings substantially. Point-level data alleviate many of the research nuisances originating from the modifiable areal unit problem. The capabilities of the interdisciplinary scientific community to employ GIS, spatial analysis and public health principles during an epidemiologic investigation warrant access to better resolution health data. In the end, it is likely to become a debate about the protection of a population's health versus the protection of a population's confidentiality. We must find the middle ground and push forward.

REFERENCES

ACS (American Cancer Society). 2009. "Breast Cancer Facts and Figures, 2009-2010." Available from http://www.cancer.org/acs /groups/ content/@nho/documents/ document/ f861009final90809pdf.pdf (accessed 20-October-2010).

ACS (American Cancer Society). 2010. "Cancer Facts and Figures 2010." Available from http://www.cancer.org/acs/groups/content/@epidemiologysurveilance/ documents/document/acspc-026238.pdf (accessed 9-November-2010).

ATSDR (Agency for Toxic Substances and Disease Registry). 1999. "Total Petroleum Hydrocarbons." Available from http://www.atsdr.cdc.gov/toxfaqs/tfacts123.pdf (accessed 11-December-2010).

Ames, B. and L. Gold. 2000. "Paracelsus to parascience: the environmental cancer distraction." *Mutation Research* 447: 3-13. Accessed 10-November-2010. http://potency.lbl.gov/pdfs/ Paracelsus.pdf.

Andrienko, G. and N. Andrienko. 1999. "Interactive maps for visual data exploration." *International Journal of Geographical Information Science* 13(4): 355-374.

Andrienko, G., N. Andrienko, and P. Gatalsky. 2000. "Mapping spatio-temporal data for exploratory analysis." *German National Research Center for Information Technology, Institute for Autonomous Intelligence Systems (AIS), Knowledge Discovery Research Group,* Sankt Augustin, Deutschland. Available from http://epp .eurostat.ec.europa.eu/ portal/page/portal/research_methodology/documents/04.pdf (accessed 10-October-2011).

Anselin, L. 1993. "The Moran Scatterplot as an ESDA Tool to Assess Local Instability in Spatial Association." Paper presented at GISDATA *Specialist Meeting on GIS and Spatial Analysis*, The Netherlands, December 1-5, 1993. Available from http://www.rri.wvu.edu/pdffiles/wp9330.pdf (accessed 05-August-2011).

Anselin, L. 1995. "Local Indicators of Spatial Association – LISA." *Geographical Analysis* 27(2): 93-115.

Belasco, A. 2010. "The Cost of Iraq, Afghanistan, and Other Global War on Terror Operations Since 9/11." Congressional Research Service. Available from http://www.fas. org/sgp/crs/natsec/RL3 3110.pdf (accessed on 4-December-2010).

Birnbaum, L. and S. Fenton. 2003. "Cancer and developmental exposure to endocrine disruptors." *Environmental Health Perspectives* 111(4): 389-394. Available from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1241417/pdf/ehp0111-000389.pdf (accessed 12-November-2010).

Bithell, J. 1995. "The choice of test for detecting raised disease risk near a point source." *Statistics in Medicine* 14(21): 2309-2322.

Bithell, J. 1999. Disease Mapping Using the Relative Risk Function Estimated from Areal Data. *Disease Mapping and Risk Assessment for Public Health* (ed.) by A. Lawson, A. Biggeri, D. Bohning, E. Lesaffre, J. Viel, and R. Bertollini. John Wiley and Sons, New York, Pp. 247-255.

Brody, J. and Rudel, R. 2003. "Environmental Pollutants and Breast Cancer." *Environmental Health Perspectives* 111(8): 1007-1019. Available from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1241551/pdf/ehp0111-001007.pdf (accessed 9-November-2010).

Brody, J., R. Rudel, K. Michels, K. Moysisch, L. Berstein, K. Attfield, and S. Gray. 2007. "Environmental pollutants, diet, physical activity, body size, and breast cancer: Where do we stand in research to identify opportunities for prevention?" *Cancer* 109(Supp 12): 2627-2634. Accessed 23-October-2010. DOI: 10.1002/ cncr.22656.

Buschbach, T., and D. Kolata. "Regional setting of Illinois Basin," in *Interior cratonic basins: American Association of Petroleum Geologists Memoir 51*, edited by Leighton, M., D. Kolata, D. Oltz, and J. Eidel, Pp. 29-55, 1990.

Campbell, N., J. Reece, L. Urry, M. Cain, S. Wasserman, P. Minorsky, and R. Jackson. *Biology*. San Francisco; Benjamin Cummings Publishing, p346, 2008.

Census Scope. 2010. "Illinois Age Distribution, 2000." Available from http://www.censusscope.org/us/s17/chart_age.html (accessed on 9-May-2012).

Cole, P. and B. Macmahon. 1969. "Oestrogen fractions during early reproductive life in the aetiology of breast cancer." *The Lancet* 293(7595): 604-606.

Cone, M. 2010. "President's Cancer Panel: Environmentally caused cancers are grossly underestimated and needlessly devastate American lives." Available from http://www.environmentalhealthnews.org/ehs/news/presidents-cancer-panel (accessed 11-December-2010).

Crisp, T., E. Clegg, R. Cooper, W. Wood, D. Anderson, K. Baetcke, J. Hoffman, M. Morrow, D. Rodier, J. Schaeffer, L. Touart, M. Zeeman, and Y. Patel. 1998. "Environmental Endocrine Disruption: An Effects Assessment and Analysis." *Environmental Health Perspectives* 106(supp 1): 11-56. Available from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1533291/pdf/envhper00536-0026.pdf (accessed 11-December-2010).

Cutler, D. "The Lifetime Costs and Benefits of Medical Technology." Harvard University. Presented October 2007. Available from http://dash.harvard.edu/bitstream/handle/1/2643640/cutler_lifetimecosts.pdf ?sequence=2.(accessed 10-December-2010).

Davis, D., D. Axelrod, L. Bailey, M. Gaynor, and A. Sasco. 1998. “Rethinking Breast Cancer Risk and the Environment: The Case for the Precautionary Principle.” *Environmental Health Perspectives* 106(9): 523-529.

DOE (Department of Energy). 2010. “US Refineries Operable Capacity.” USEPA Information Administration. Available from http://www.eia.doe.gov/neic/ rankings/refineries.htm (accessed 27-November-2010).

Dolecek T., T. Shen, and J. Snodgrass. 2003. “Illinois County Cancer Statistics Review, Incidence, 1996-2000.” Epidemiologic Report Series 03:4. Springfield, Ill.: Illinois Department of Public Health, 2003. Available from http://www.idph.state. il.us/cancer/ pdf/Cty9600.pdf (accessed 27-November-2010).

E-Health MD. 2010. "What Causes Breast Cancer." Available from http://www.eheal thmd.com/library/breastcancer/brc_causes.html (accessed 9-November-2010).

E-Medicine. 2010. “Breast Cancer Causes.” Available from http://www.emedicine health.com/breast_cancer/page2_em.htm (accessed 9-November-2010).

EPA (Environmental Protection Agency). 2007. “EPA, Federal Register.” EPA Federal Register 72(116). Available from http://www.epa.gov/scipoly/oscpendo/pubs/ draft_list_frn_061807.pdf (accessed 8-November-2010).

EPA (Environmental Protection Agency). 2009. “News Releases – Pesticides and Toxic Chemicals.” Available from http://yosemite.epa.gov/opa/admpress.nsf/eeff e922a687433c85257359003f5340/f34f79b60fcdf4eb852575690068995b!OpenDocument (accessed 4-December-2009).

EPA (Environmental Protection Agency). 2010a. “Safe Drinking Water Act.” Available from http://water.epa.gov/drink/standardsriskmanagement.cfm (accessed 27-November-2010).

EPA (Environmental Protection Agency). 2010b. "Drinking Water Contaminants.” Available from http://water.epa.gov/drink/contaminants/index.cfm (accessed 23-October-2010).

EPA (Environmental Protection Agency). 2010c. “Endocrine Disruptors.” Available from http://www.epa.gov/scipoly/oscpendo/pubs/edspover view/primer.htm (accessed 8-November-2010).

EPA (Environmental Protection Agency). 2010d. “Radioactive Wastes from Oil and Gas Drilling.” Available from http://www.epa.gov/radtown/ drilling-waste.html (accessed 11-December-2010).

EPA (Environmental Protection Agency). 2012a. “Waste Types.” Available from http://www.epa.gov/osw/hazard/wastetypes/index.htm (accessed on 7-May-2012).

EPA (Environmental Protection Agency). 2012b. "EPA Geospatial Data Access Project." Last updated 2-January-2012. Available from http://www.epa.gov /enviro/geo_data.html (accessed on 6-August-2011).

EPA (Environmental Protection Agency). 2012c. "History of UIC Program – Injection Well Time Line." Available from http://water.epa.gov/type/groundwater/uic/ history.cfm (accessed on 05/06/2012).

EPA (Environmental Protection Agency). 2012d. "Hazardous Substance and Hazardous Waste." Available from http://www.epa.gov/superfund/students/clas_act/haz-ed/ff_01.htm (accessed on 5/05/2012).

Edsall, R., G. Andrienko, N. Andrienko, and B. Buttenfield. Interactive Maps for Exploring *Spatial Data*. Marguerite Madden (ed.) ASPRS Manual of GIS, 2008.

Ejaz, S., W. Akram, C. Lim, L. Woong, J. Jong, I. Hussain. 2004. "Endocrine disrupting pesticides: A leading cause of cancer among rural people in Pakistan." *Experimental Oncology* 26(2): 98-105.

Ezra, S., S. Feinstein, A. Yakirevich, E. Adar, and I. Bilkis. 2006. "Retardation of organo-bromides in a fractured chalk aquitard." *Journal of Contaminant Hydrology* 86: 195-214. Accessed 5-June-2011. DOI:10.1016/j.jcon hyd.2006.02.016.

Gelman, A. 2004. "Exploratory Data Analysis for Complex Models." *Journal of Computational and Graphical Statistics* 13(4): 755–779. DOI: 10.1198/106186004X11435.

Gelman, A. 2011. "Why tables really are much better than graphs." *Journal of Computational and Graphical Statistics* 20(1): 3-7.

Getis, A. and K. Ord. 1995. "Local Spatial Autocorrelation Statistics: Distributional Issues and an Application." *Geographical Analysis* 27(4): 286-306. DOI: 10.1111/j.1538-4632.1995.tb00912.x.

Goldsmith, D. 1987. "Calculating cancer latency using data from a nested case-control study of prostatic cancer." *Journal of Chronic Disease* 40(suppl 2): 119-123.

Gray, J. 2010. "State of the Evidence 2010: The Connection Between Breast Cancer and the Environment." Breast Cancer Fund, 2010. Available from http://www.bre astcancerfund.org/assets/pdfs/ publications/state-of-the-evidence-2010.pdf (accessed 11-December-2010).

Greenblatt, M., W. Bennett, M. Hollstein, and C. Harris. 1994. "Mutations in the p53 Tumor Suppressor Gene: Clues to Cancer Etiology and Molecular Pathogenesist." *Cancer Research* 54: 4855-4878. Available from http://cancerres.aacrjournals.org/content/ 54/18/4855.full.pdf. (accessed 11-November-2010).

Greenlee, R., T. Murray, S. Bolden, and P. Wingo. 2000. "Cancer Statistics, 2000." *CA Cancer Journal for Clinicians* 50: 7-33. Available from http://www.tcsg.org/ tobacco/CancerStat2000.pdf (accessed 27-April-2011).

Groschen, G., M. Harris, R. King, P. Terrio, and K. Warner. 2000. "Water Quality in the Lower Illinois River Basin, Illinois, 1995-1998." US Geological Survey Cicular 1209. Available from http://pubs.water. usgs.gov/circ1209/ (accessed 5-December-2010).

Guajardo, O. and T. Oyana. 2009. "A Critical Assessment of Geographic Clusters of Breast and Lung Cancer Incidences among Residents Living near the Tittabawassee and Saginaw Rivers,Michigan, USA." *Journal of Environmental and Public Health* 2009: 1-16. Accessed 10-December-2010. DOI:10.1155/ 2009/316249.

Hosmer, D. and S. Lemeshow. *Applied Logistic Regression*, 2nd Ed. John Wiley and Sons, Inc: New York, NY, 2000.

Huggins, F., L. Seidu, N. Shah, G. Huffman, R. Honaker, J. Kyger, B. Higgins, J. Robertson, S. Pal, and M. Seehra. 2009. "Elemental modes of occurrence in an Illinois #6 coal and fractions prepared by physical separation techniques at a coal preparation plant." *International Journal of Coal Geology* 78: 65-76. Accessed 3-December-2010. DOI:10.1016/j.coal.2008.10.002.

IDNR (Illinois Department of Natural Resources). 2010. "Oil and Gas Facts." Available from http://dnr.state.il.us/mines/dog/facts.htm (accessed 10-December-2010).

IDPH (Illinois Department of Public Health). 2008. "Radium in drinking water." Available from http://www.idph.state.il.us/envhealth/factsheets/ radium.htm (accessed 8-December-2010).

IEPA (Illinois Environmental Protection Agency). 2009. "ConocoPhillips Wood River Refinery Located in Wood River, Illinois, Madison County NPDES Permit Responsiveness Summary." Available from http://www.epa state.il.us/public-notices/2007/conoco-phillips-wood-river/responsiveness-summary-il0000205.pdf (accessed 27-November-2010).

IEPA (Illinois Environmental Protection Agency). 2010. "Illinois EPA's Ash Impoundment Strategy Progress Report." Available from http://www.epa.state.il. us/water/groundwater/publications/ash-impoundment-progress.pdf (accessed 20-November-2010).

ISGS (Illinois State Geological Survey). 2010a. "Oil Fields in Illinois." Available from http://www.isgs. illinois.edu/maps-data-pub/publications/geobits/geobit9.shtml (accessed 15-November-2010).

ISGS (Illinois State Geological Survey). 2010b. “Illinois Coal Resource Shapefiles.” Available from http://www.isgs.illinois.edu/maps-data-pub/coal-maps/coalshapefiles.shtml (accessed 10-November-2010).

ISGS (Illinois State Geological Survey) 2010c. "Trace Elements Maps." Available from http://www.isgs.illinois.edu/maps-data-pub/coal-maps/trace-elements.shtml (accessed 20-November-2010).

ISGS (Illinois State Geological Survey). 2010d. “Depositional History of the Pennsylvanian Rocks in Illinois.” Available from http://www.isgs.illinois.edu/ maps-data-pub/publications/geonotes/geonote2.shtml (accessed 4-December-2010).

ISGS (Illinois State Geological Survey). 2010e. "Location Points from the ISGS Wells and Borings Database." Available from http://www.isgs.illinois. edu/nsdihome/webdocs/st-geolb.html (accessed 19-November-2010).

ISGS (Illinois State Geological Survey). 2010f. “Coal – Illinois Black Treasure.” Available from http://www.isgs.illinois. edu/maps-data-pub/publications/geobits/ geobit12.shtml (accessed 5-December-2010).

ISWS (Illinois State Water Survey). 2009a. “Arsenic in Illinois Groundwater.” Available from http://www.isws.illinois.edu/gws/arsenic/ilsources.asp (accessed 27-November-2010).

ISWS (Illinois State Water Survey). 2009b. "Including the Mahomet Teays Aquifer System in a National Groundwater Monitoring Network 2010." Available from http://acwi.gov/sogw/pubs/tr/Il-id_soi_mahomet_teays_ aquifer.pdf (accessed 10-November-2010).

ISWS (Illinois State Water Survey). 2009c. "Mahomet Aquifer." Available from http://www.isws.illinois.edu/gws/mahomet.asp (accessed 19-November-2010).

Inman, R. 2010. "States in Fiscal Distress." *Federal Reserve Bank of St. Louis Regional Economic Development* 6(1): 65-80. Available from http://research. stlouisfed.org/publications/red/2010/01/Inman.pdf (accessed 11-December-2010).

Jacquez, G. 2011. “The small numbers problem-Part 2.” Available from http://www.biomedware.com/blog/2011/the-small-numbers-problem-part-2-using-persistence-in-spatial-time-series-as-a-diagnostic-for-extreme-rates-in-small-areas/ (accessed on 15-October-2011).

Jakhrani, A., S. Samo, and I. Nizamani. 2009. “Impact of Wastewater Effluents on Physico-Chemical Properties of Groundwater.” *Sindh Univ. Res. Journal* 41(1): 75-82. Available from http://usindh.edu.pk/surj/volume_41_01/a.q.ja khr%2012.pdf (accessed 5-June-2011).

Jemal, A., R. Siegel, J. Xu, and E. Ward. 2010. "Canster Statistics 2010." CA: A Cancer Journal for Clinicians 60: 277-300. Accessed 15-October-2010. DOI: 10.3322/caac.20073.

Jüntgen, H. and J. Klein. 1977. "Purification of Wastewater from Coking and Coal Gasification Plants Using Activated Carbon." *Energy Sources* 2(4): 67-84. Available from http://www.anl.gov/PCS/acsfuel/preprint% 20archive/Files/19_5_ATLANTIC%20CITY_09-74_0067.pdf (accessed 20-November-2010).

Jurek, A., G. Maldonado, S. Greenland, and T. Church. 2006. "Exposure measurement error is frequently ignored when interpreting epidemiological study results." European Journal of Epidemiology 21: 871-876. Accessed 16-November-2010. DOI: 10.1007/s10654-006-9083-0.

KFF (Kaiser Family Foundation). 2010. "How Changes in Medical Technology Affect Health Care Costs, March 2007." Available from http://www.kff.org/insurance/snapshot/chcm030807oth.cfm (accessed 10-December-2010).

Kalyoncu, R. 1997. "Coal Combustion Products." American Coal Ash Association. Available from http://minerals.usgs.gov/minerals/pubs/commodity/coal/874497.pdf (accessed 20-November-2010).

Kelly, W. 2008. "Radium and Barium in the Ironton-Galesville Bedrock Aquifer in Northeastern Illinois Final Report." Available from http://mtac.isws.illinois.edu/mtacdocs/pubs/MTACTR08-01.pdf (accessed 19-November-2010).

Khan, K., M. Suidan, and W. Cross. 1981. "Anaerobic activated carbon filter for the treatment of phenol-bearing wastewater." Water Pollution Control Federation 53(10): 1519-1532. Available from http://www.jstor.org/stable/25041533 (accessed 24-November-2010).

Korte, N. and Q. Fernando.1991. "A review of arsenic (III) in groundwater." *Critical Review of Environmental Control* 21: 1-11. Accessed 19-November-2010. DOI: 10.1080/10643389109388408.

Krieger, N. 1989. "Exposure, susceptibility and breast cancer risk: a hypothesis regarding exogenous carcinogens, breast tissue development, and social gradients, including black~white differences, in breast cancer incidence." *Breast Cancer Research and Treatment* 13: 205-223.

Kruglyak, L., M. Daly, M. Reeve-Daly, and E. Lander. 1996. "Parametric and nonparametric linkage analysis: a unified multipoint approach." *American Journal of Human Genetics* 58: 1347-1363.

Lander, E. and P. Green. 1987. "Construction of multilocus genetic maps in humans." *Proceedings of National Academy of Sciences USA* 84: 2363-2367.

Lawson, A. *Score tests for detection of spatial trend in morbidity data*. Dundee Institute of Technology, Dundee, 1989.

Lawson, A. *Statistical Methods in Spatial Epidemiology*. John Wiley and Sons: West Sussex, England, 2006.

Lee, J and D. Wong. 2005. *Statistical Analysis with ArcView GIS*. John Wiley & Sons, New York, USA.

Lewicki, P. and T. Hill. 2007. *Statistics: Methods and Applications*. StatSoft, Tulsa, OK. Available from http://books.google.com/books?hl=en&lr=&id=TI2TGjeilMAC&oi=fnd&pg=PR15&dq=lewicki+and+hill+statistics:+methods+and+applications&ots=PwmCxqEN2r&sig=ZRLTYhzr-YIjr27D8-gYtfUWnbM#v=onepage&q&f=false (accessed 7-December-2010).

Lichtenstein, P., N. Holm, P. Verkasalo, A. Iliadou, J. Kaprio, M. Koskenvuo, E. Pukkala, A. Skytthe, K. Hemminki. 2000. "Environmental and heritable factors in the causation of cancer." *The New England Journal of Medicine* 343(2): 78-85. Available from ftp://stat-www.berkeley.edu/pub/users/terry/Classes/s246.2002/Week4/lichtenstein.pdf (accessed 20-October-2010).

Locatelli, I., P. Lichtenstein, and A. Yashin. 2004. "The heritability of breast cancer: a Bayesian correlated frailty model applied to Swedish twins data." *Twin Research* 7(2): 182-191.

MNRG (Midwest Natural Resources Group). 2005. "Illinois River Maps." Available from http://www.mnrg.gov/accomplishments/illinois-river-maps.htm (accessed 9-December-2010).

MacEachren, A. 1982. "Map Complexity: Comparison and Measurement." *The American Cartographer* 9(1): 31-46.

Meade, M. and M. Emch. *Medical Geography*, 3rd Ed. The Guilford Press, New York, NY, 2010.

Mehnert, E., K. Hackley, T. Larson, S. Panno, and A. Pugin. 2004. "The Mahomet Aquifer: Recent Advances in our Knowledge." Available from http://www.mahometaquifer consortium.org/ofs2004-16.pdf (accessed 21-November-2010).

Miller, B., E. Feuer, and B. Hankey. 1994. "The significance of the rising incidence of breast cancer in the United States." *Important Advances in Oncology* 12: 193-207.

Moore, D. and T. Carpenter. 1999. "Spatial Analytical Methods and Geographic Information Systems: Use in Health Research and Epidemiology." *Epidemiologic Reviews* 21(2): 143-161.

Moore, J. N. 1994. "Contaminant mobilization resulting from redox pumping in a metal-contaminated river reservoir system." *Environmental Chemistry of Lakes and Reservoirs*: 451-471. Washington D.C., American Chemical Association.

Morrow, W. 1999. "Volatile Organic Compounds in Ground Water of the Lower Illinois River Basin." United States Geological Survey: Water Investigations Report 99-4229. Available from http://il.water.usgs.gov/proj/ lirb/pubs/pdfs/voc.pdf. (accessed 20-October-2010).

NASS (National Agricultural Statistics Service). 2012. "Land Cover of Illinois, 1999-2000: TM and ETM+ classification from Landsat 5 and Landsat 7." Available from http://www.agr.state.il.us/gis/landcover99-00.html (accessed on 10-October-2012).

NAWQA. 1994. "National Water Quality Assessment Program-The Lower Illinois River Basin, NAWQA Fact Sheet." FS 94-018. Available from http://il.water.usgs.gov/ proj/lirb/pubs/pdfs/fctsheet.pdf. (accessed 7-December-2010).

NCI (National Cancer Institute). 2010a. "Cancer Trends Progress Report – 2009/2010." Available from http://progressreport.cancer.gov/doc_detail.asp?pid=1&did= 2007&chid=75&coid=726&mid (accessed 3-December-2010).

NCI (National Cancer Institute). 2010b. "State Cancer Profiles." Available from http://statecancerprofiles.cancer.gov/cgi-bin/quickprofiles/profile.pl?17&055 (accessed 9-Nov-2010).

NCI (National Cancer Institute). 2010c. "Breast Cancer Prevention." Available from http://www.cancer.gov/cancertopics/pdq/prevention/breast/HealthProfessional (accessed 15-November-2010).

NEIS (Nuclear Energy Information Service). 2010. "Nuclear Energy Information Service." Available from http://www.neis.org/Content/Nuclear_Illinois.shtml (accessed on 14-March-2012).

NIH (National Institute of Health). 2002. "Nitromethane: Report on Carcinogens." Available from http://ntp.niehs.nih.gov/ntp/newhomeroc/roc11/nmpub.pdf (accessed on 15-March-2012).

NIH (National Institute of Health). 2011. "Nitromethan: Report on Carcinogens." Available from http://ntp.niehs.nih.gov/ntp/roc/twelfth/profiles/Nitromethane.pdf (accessed on 15-March-2012).

NYDPH (New York Department of Public Health). 2006. "About Age Adjusted Rates." Available from http://www.health.ny.gov/statistics/cancer/registry/age.htm (accessed on 5-January-2012).

Nordling, C. 1953. "A New Theory on the Cancer-inducing Mechanism." *British Journal of Cancer* 7(1): 68-72.

Oyana, T. and J. Lwebuga-Mukasa. 2004. "Spatial relationships among asthma prevalence, health care utilization, and pollution sources in neighborhoods of Buffalo, New York." *Journal of Environmental Health* 66(8):25-37.

Oyana, T. and F. Margai. 2010. "Spatial Patterns and Health Disparities in Pediatric Lead Exposure in Chicago: Characteristics and Profiles of High-Risk Neighborhoods." *The Professional Geographer* 62(1):46-65. DOI: 10.1080/00330120903375894.

PCP (President's Cancer Panel). 2010. "Reducing Environmental Risk: What We Can Do Now." Available from http://deainfo.nci.nih.gov/advisory/pcp/annualReports/ pcp08-09rpt/PCP_Report_08-09_508.pdf (accessed 15-November-2010).

Parkin, M., F. Bray, J. Ferlay, and P. Pisani. 2005. "Global Cancer Statistics." *CA A Cancer Journal for Clinicians* 55: 74-108. Available from http://caonline. amcancersoc.org/cgi/reprint/55/2/74 (accessed 5-November-2010).

Peduzzi, P. N., J. Concato, E. Kemper, T. R. Holford, and A. Feinstein. 1996. "A simulation study of the number of events per variable in logistic regression." *Journal of Clinical Epidemiology* 99: 1373-1379.

Pincus-Nielsen, M., J. Gordon, and C. Moseley. "Monitoring the American Reinvestment and Recovery Act in the 11 Western States." White Paper. Ecosystem Workforce Program, University of Oregon. Presented spring 2010. Available from https://scholarsbank.uoregon.edu/xmlui/bitstream/handle/1794/10779/ BP_24.pdf?sequence=1 (accessed 9-December-2010).

Poulsen, J. and A. French. 2012. "Discriminant Function Analysis." Available from http://userwww.sfsu.edu/~efc/classes/biol710/discrim/discrim.pdf (accessed 29-April-2012).

Ramankutty, Navin and Foley, Jonathan A. 1999. "Estimating historical changes in land cover: North American croplands from 1850 to 1992." *Global Ecology and Biogeography* Vol 8: 381-396.

Rothman K. and S. Greenland. *Modern epidemiology,* 2nd Ed. Lippincott–Raven, Philadelphia, 1998.

Rudel, R., K. Attfield, and J. Schifano. 2007. "Chemicals causing mammary gland tumors in animals signal new directions for epidemiological chemicals testing, and risk assessment for breast cancer prevention." *Cancer*, 109 (Suppl 12): 2635-2666. Accessed 25-October-2010. DOI: 10.1002/cncr.22653.

SAS (SAS Institute 9.2). 2012a. "The PROC GLIMMIX Statement." Available from http://support.sas.com/documentation/cdl/en/statug/63033/HTML/ default/viewer. htm#statug_glimmix_a0000001405.htm (accessed on 2-May-2012).

SAS (SAS Institute 9.2). 2012b. "Repeated Statement." Available from http://support.sas.com/ documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_mixed_sect019.ht m (accessed on 5-June-2012).

SEER (Survey Epidemiology and End Results). 2010a. "SEER Cancer Statistics Review, 1975-2007." Available from http://seer.cancer.gov/csr/1975_2007/ browse_csr.php (accessed 9-December-2010).

SEER (Survey Epidemiology and End Results). 2010b. "SEER Cancer Statistics Review, 1975-2000." Available from http://seer.cancer.gov/csr/1975_2000/ results_merged/sect_04_breast.pdf (accessed 27-November-2010).

SEER (Surveillance Epidemiology and End Results). 2011. "SEER Tutorials: Calculating Age Adjusted Rates." Available from http://seer.cancer.gov/seerstat/ tutorials/aarates/definition.html (accessed on 5-January-2011).

Slocum, T., R. McMaster, F. Kessler, and H. Howard. *Thematic Cartography and Geovisualization*. Prentice Hall, New Jersey, 2009.

Stewart, J., A. Lemley, S. Hogan, R. Weismiller, and A. Hornsby. "Drinking Water Standards, SL159." University of Florida Press, 2001. Available from http://edis.ifas.ufl.edu/pdffiles/SS/ SS29700.pdf (accessed 28-November-2010).

Struewing, J., P. Hartge, S. Wacholder, S. Baker, M. Berlin, M. McAdams, M. Timmerman, L. Brody, and M. Tucker. 1997. "The Risk of Cancer Associate with Specific Mutations of BRCA1 and BRCA2 among Ashkenazi Jews." *The New England Journal of Medicine* 336: 1401-1408.

Tobler, W. "Cellular Geography." In *Philosophy in geography* (ed.) S. Gale and G. Olsson, 379-386. Dordrecht, The Netherlands, 1979.

Treworgy, C. and R. Jacobson. 1986. "Paleoenvironments and distribution of low sulfur coal in Illinois." IX-ICC 4: 349-359. Available from http://www.isgs.illinois.edu/maps-data-pub/publications/pdf-files/reprint1986e.pdf (accessed 5-December-2010).

Tukey, J. "Some Graphic and Semigraphic Displays." In *Statistical Papers in Honor of George W. Snedecor* (ed.) T. Bancroft, Iowa State University Press, 1972.

US Census Bureau. 2010a. "Profile of General Demographic Characteristics: 2000." Available from http://factfinder.census.gov/servlet/QTTable?_bm=y&-geo_id=04000US17&-qr_name=DEC_2000_ SF1_U_DP1&-s_name=DEC_ 2000_ SF1_U (accessed 28-November-2010).

US Census Bureau. 2010b. "National and State Population Estimates." Available from http://www.census.gov/popest/ (accessed 13-December-2010).

US Census Bureau. 2010c. "Cartographic Boundary Files." Available from http://www.census.gov/geo/www/cob/z52000.html#shp (accessed on 10-November-2010).

US Census Bureau. 2010d. "Download Center." Available from http://factfinder. census.gov/servlet/DownloadDatasetServlet?_lang=en (accessed on 5-May-2011).

USEIA (United States Energy Information Association). 2009. "Map of Illinois."Available from http://205.254.135.7/state/state-energy-profiles.cfm?sid=IL (accessed on 24-March-2012).

USGS (United States Geologic Survey). 1997. "A rebirth of the Illinois Basin" Available from http://energy.usgs.gov/factsheets/Illinois/illinois.basin.html (accessed 20-November-2010).

USGS (United States Geologic Survey). 1999. "Radium in ground water from public-supply aquifers in northern Illinois." Available from http://il.water.usgs.gov /proj/gwstudies/radium/ (accessed 8-December-2010).

USGS (United States Geologic Survey). 2006. "Arsenic in Coal." Available from http://pubs.usgs.gov/fs/2005/3152/ (accessed 20-November-2010).

Union of Concerned Scientists. 2009. "Clean Energy- How Coal Works." Available from http://www.ucsusa.org/clean_energy/technology_and_impacts/energy_techno logies/how-coal-works.html#vi_Bonskowski_R_et_al_2006_Coal_producti (accessed 21-November-2010).

Voldner, E. and Y. Li. 1995. "Global usage of selected persistent organochlorines." *The Science of the Total Environment* 160/16: 201-210.

Vrijheid, M. 2000. "Health effects of residence near hazardous waste landfill sites: a review of epidemiologic literature." *Environmental Health Perspectives* 108(Supp 1): 101-112. Available from http://www.ncbi.nlm.nih.gov/pmc/ articles/PMC1637771/ ?page=1 (accessed 5-March-2012).

WHO (World Health Organization). 2010. "World Health Statistics 2009." Available from http://www.who.int/whosis/whostat/ 2009/en/index.html (accessed 13-December-2010).

Waller, L., B. Turnbull, L. Clark, and P. Nasca. 1992. "Chronic disease surveillance and testing of clustering of disease and exposure: Application to leukemia incidence and TCE-contaminated dumpsites in upstate New York." *Environmetrics* 3(3): 281-300.

Waller, L. and C. Gotway. *Applied Spatial Statistics for Public Health Data*. John Wiley and Sons, New York, 2004.

Wang, F. 2004. "Spatial clusters of cancer in Illinois 1986-2000." *Journal of Medical Systems* 28(3): 237-256. Accessed 19-October-2010. DOI: 10.1023/B:JOMS. 0000032842.78643.38.

Warner, K. 2001. "Arsenic in glacial drift aquifers and the implication for drinking water-Lower Illinois River Basin." *Ground Water* 39(3): 433-442.

Warner, K., A. Martine, and T. Arnold. 2003. "Arsenic in Illinois Ground Water-Community and Private Supplies." Water-Resources Investigations Report 03-4103. Available from http://il.water.usgs.gov/pubs /wrir03_4103.pdf (accessed 20-November-2010).

Welch, A., M. Lico, and J. Hughes. 1988. "Arsenic in ground water of the Western United States." *Ground Water* 26: 333-347. Accessed 29-October-2010. DOI: 10.1111/ j.1745-6584.1988.tb00397.x.

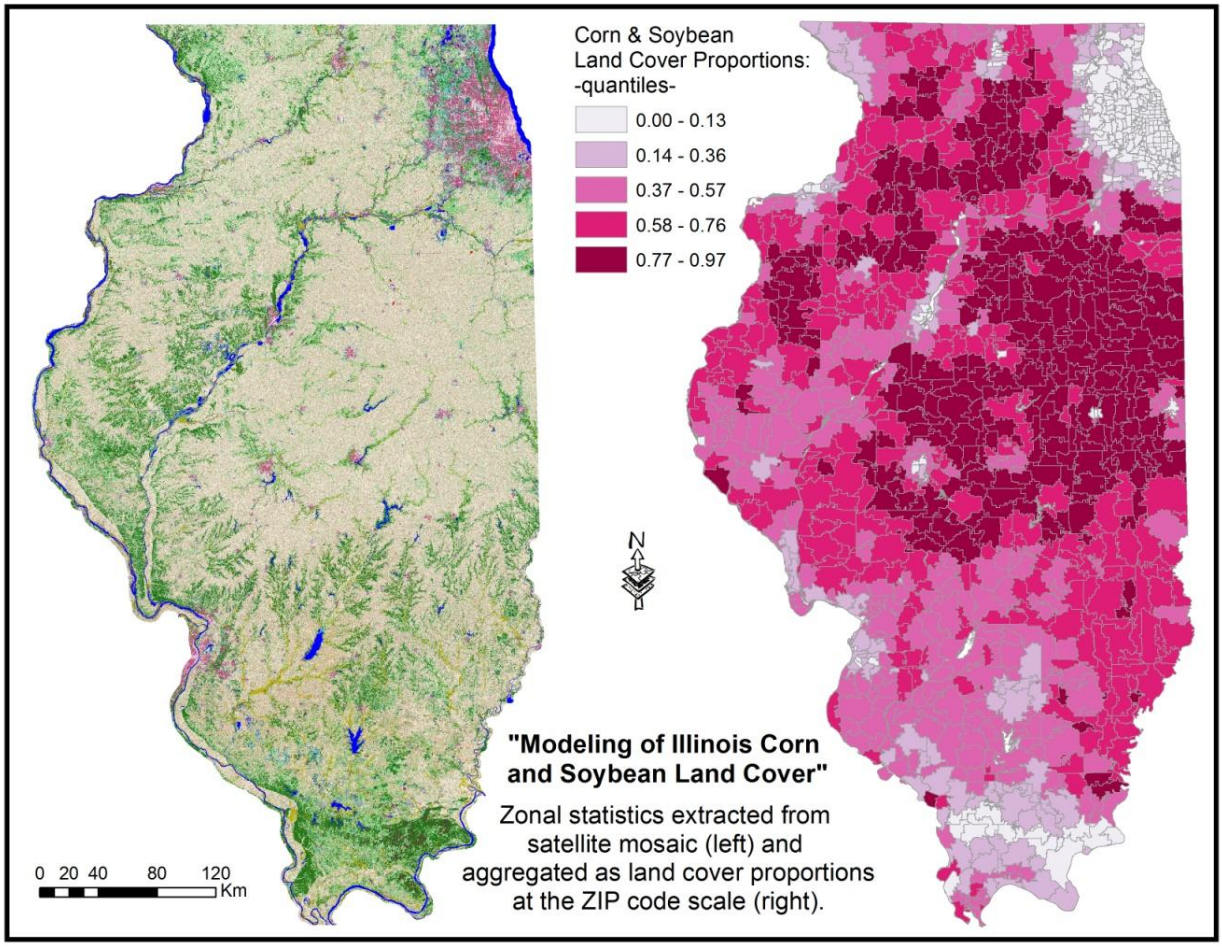
Welch, A., D. Westjohn, D. Helsel, and R. Wanty. 2000. "Arsenic in ground water of the United States: Occurrence and Geochemistry." *Ground Water* 38(4): 589-604. Available from http://water.usgs.gov/nawqa/trace/pubs/g w_v38n4/ (accessed 29-October-2010).

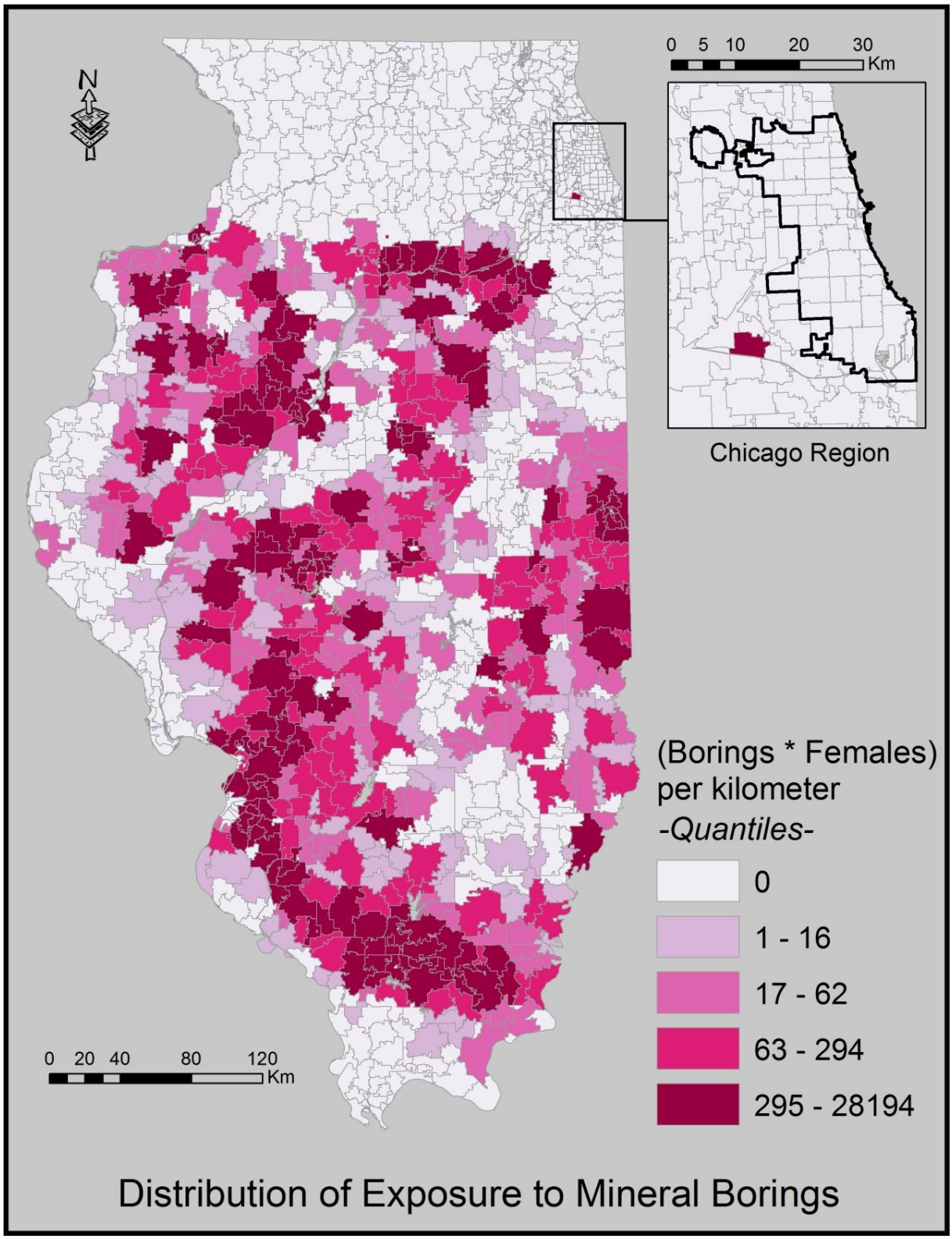
Wolff, M. and P. Toniolo. 1995. "Environmental Organochlorine Exposure as a Potential Etiologic Factor in Breast Cancer." *Environmental Health Perspectives* 103(Supp 7): 141-145. Available from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC151 8872/pdf/envhper00367-0137.pdf (accessed 28-November-2010).

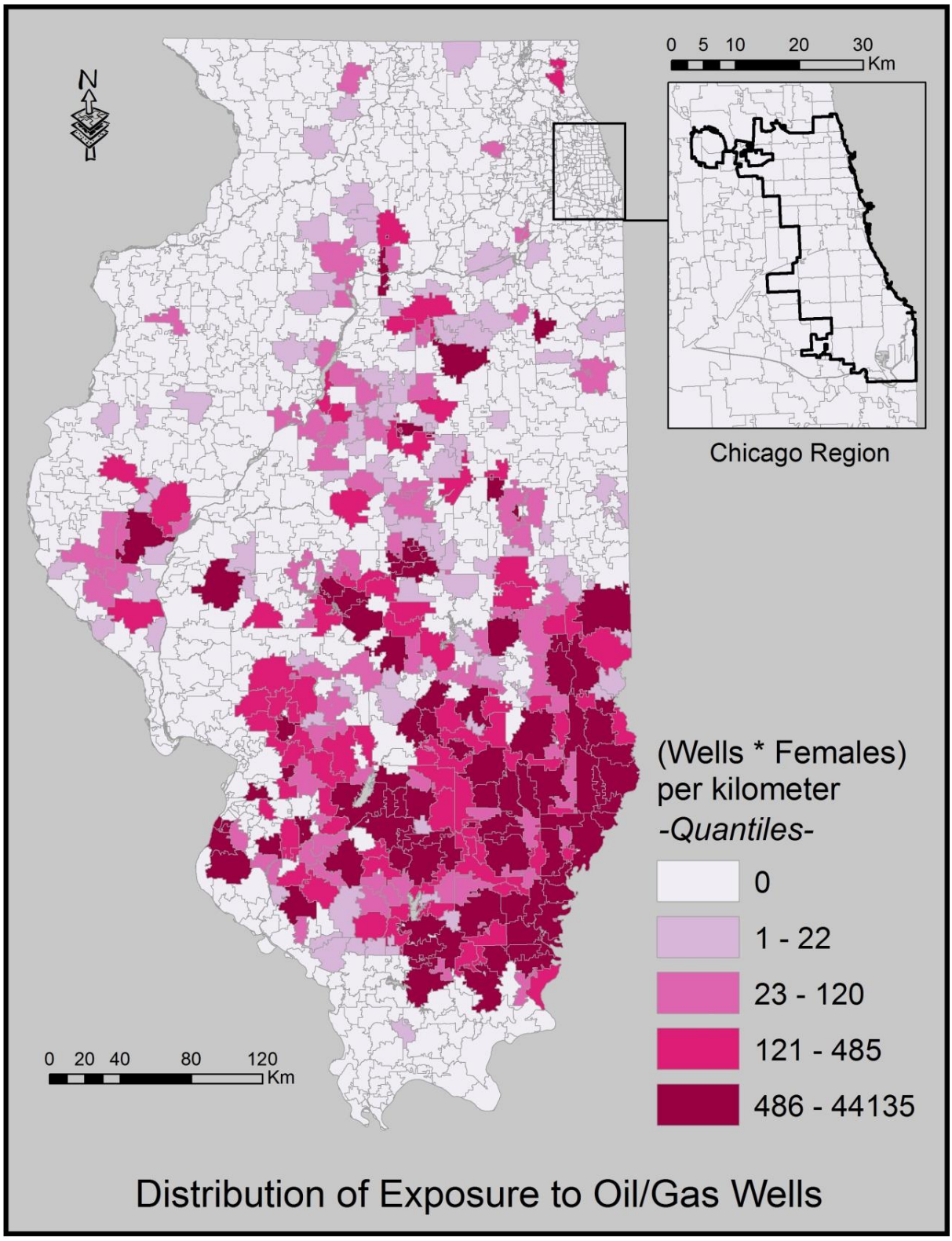
Worrall, M. and I. Zuber. 2010. "Control VOC's in Refinery Wastewater." Process Optimization Conference, Houston, TX. Available from http://www.amcec.com/ case3.html (accessed 27-November-2010).

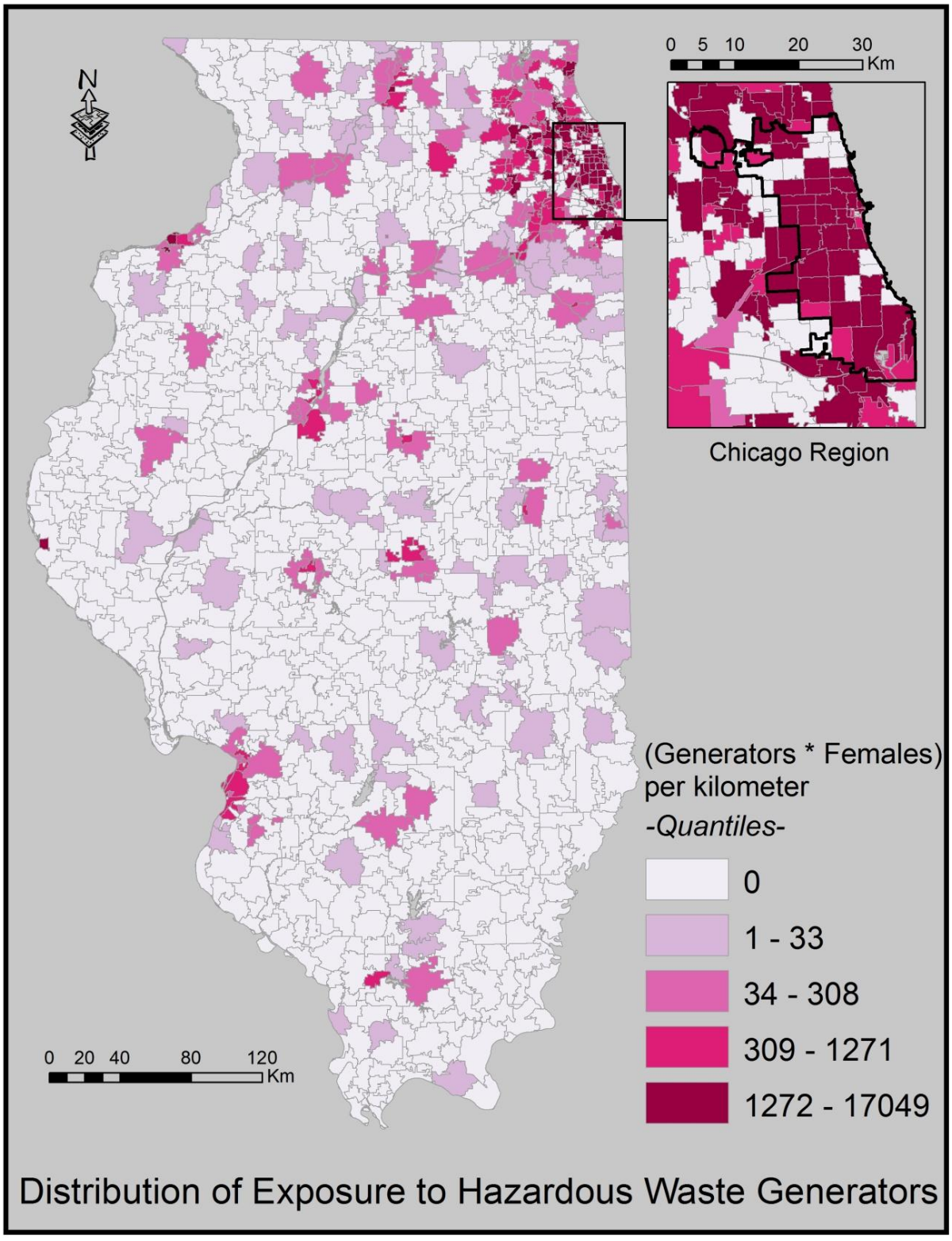
APPENDICIES

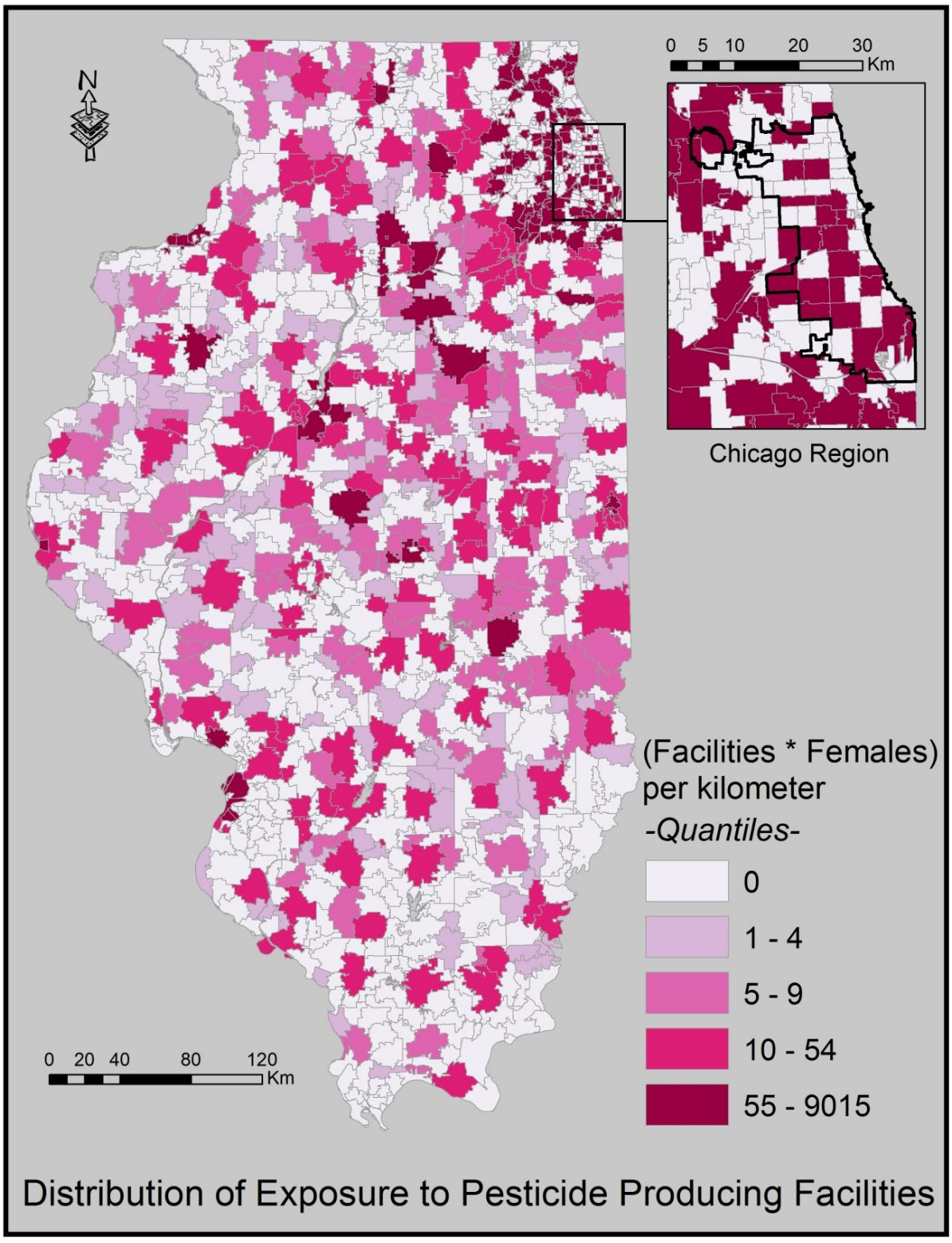
APPENDIX A

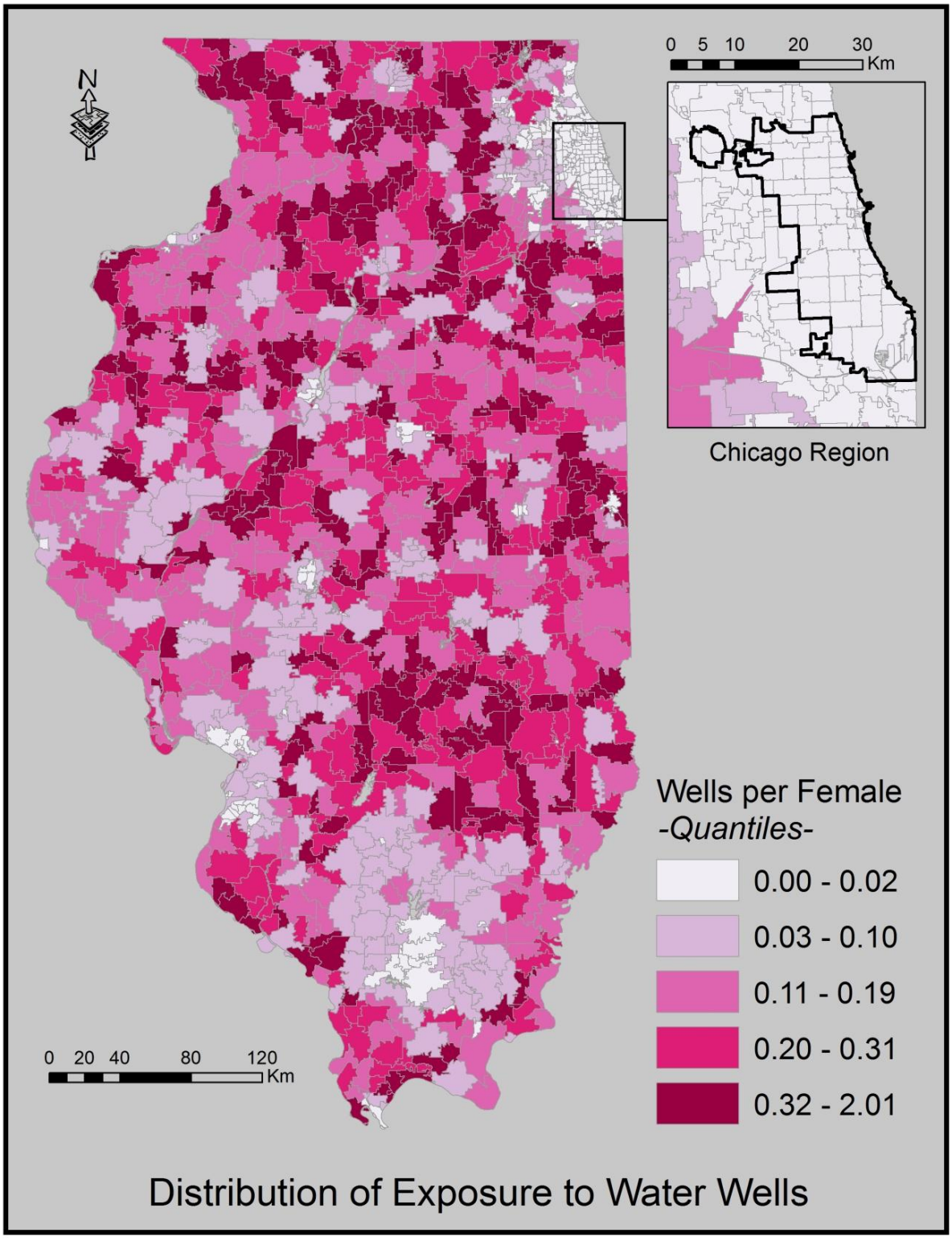

Corn & Soybean
Land Cover Proportions:
-quantiles-
0.00 - 0.13
0.14 - 0.36
0.37 - 0.57
0.58 - 0.76
0.77 - 0.97
N
0 20 40 80 120
Km
"Modeling of Illinois Corn and Soybean Land Cover"
Zonal statistics extracted from satellite mosaic (left) and aggregated as land cover proportions at the ZIP code scale (right).

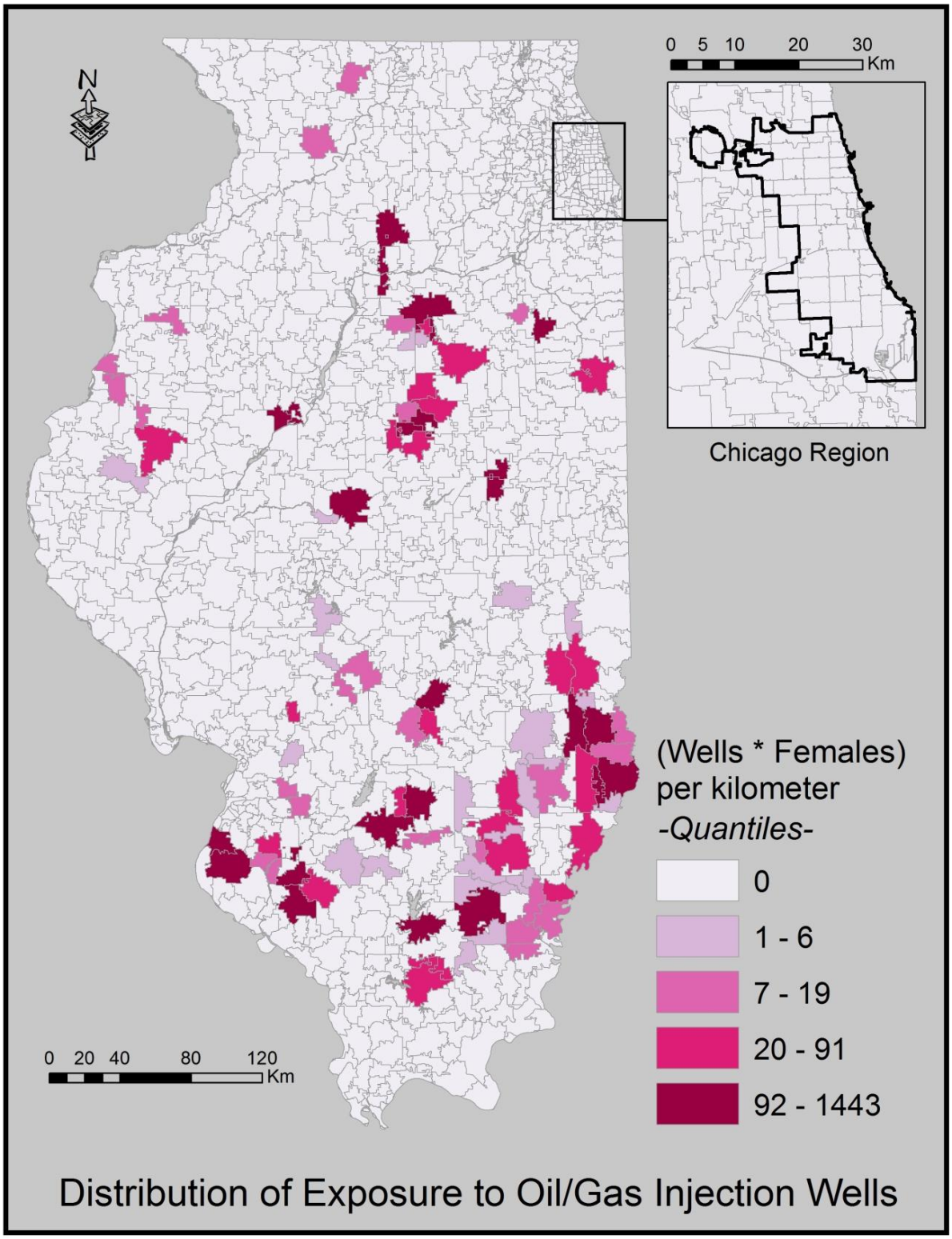

N
0 5 10 20 30
Km
Chicago Region
(Borings * Females)
per kilometer
-Quantiles-
0
1 - 16
17 - 62
63 - 294
295 - 28194
0 20 40 80 120
Km
Distribution of Exposure to Mineral Borings

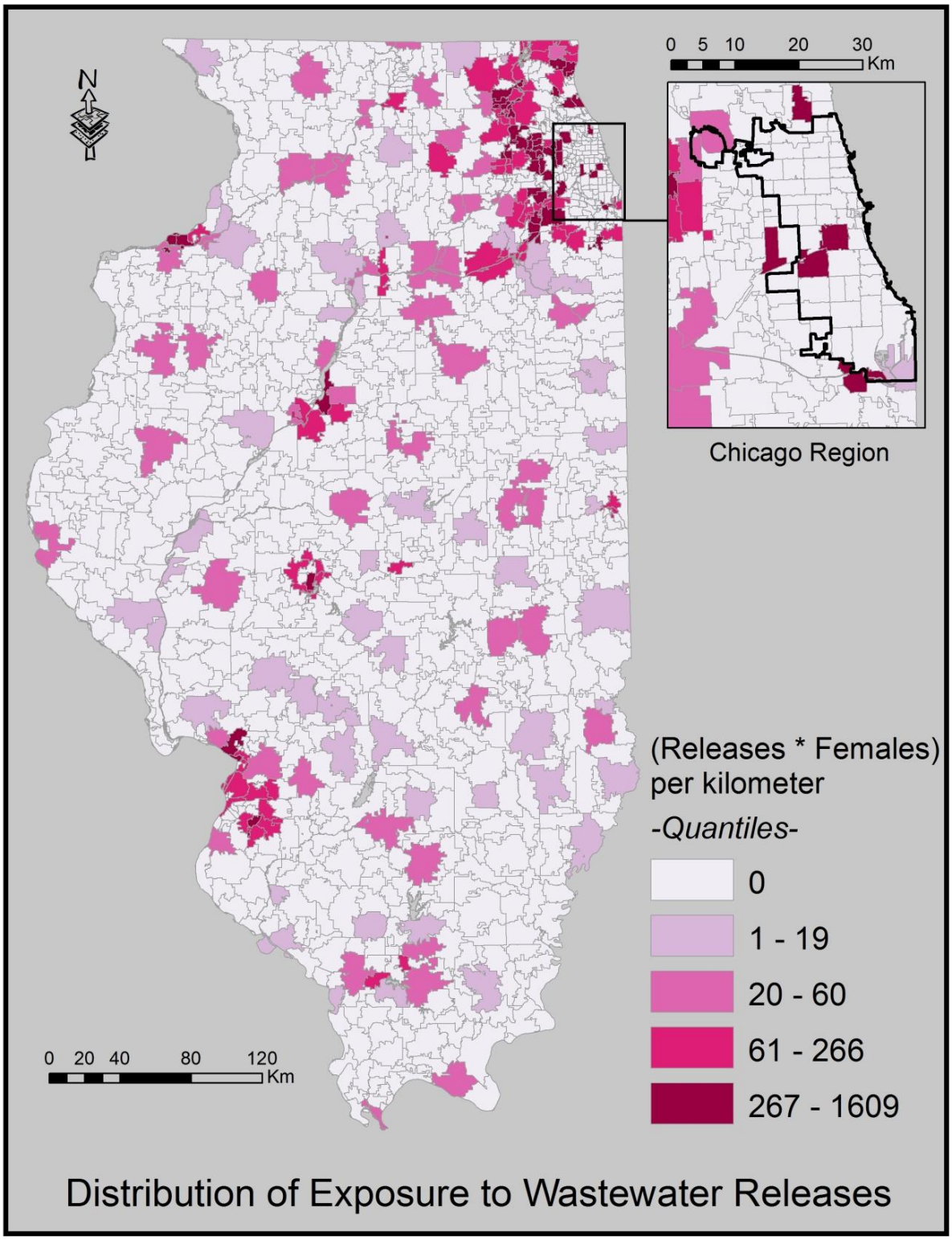

N
0 5 10 20 30
Km
Chicago Region
(Wells * Females)
per kilometer
-Quantiles-
0
1 - 22
23 - 120
121 - 485
486 - 44135
0 20 40 80 120
Km
Distribution of Exposure to Oil/Gas Wells

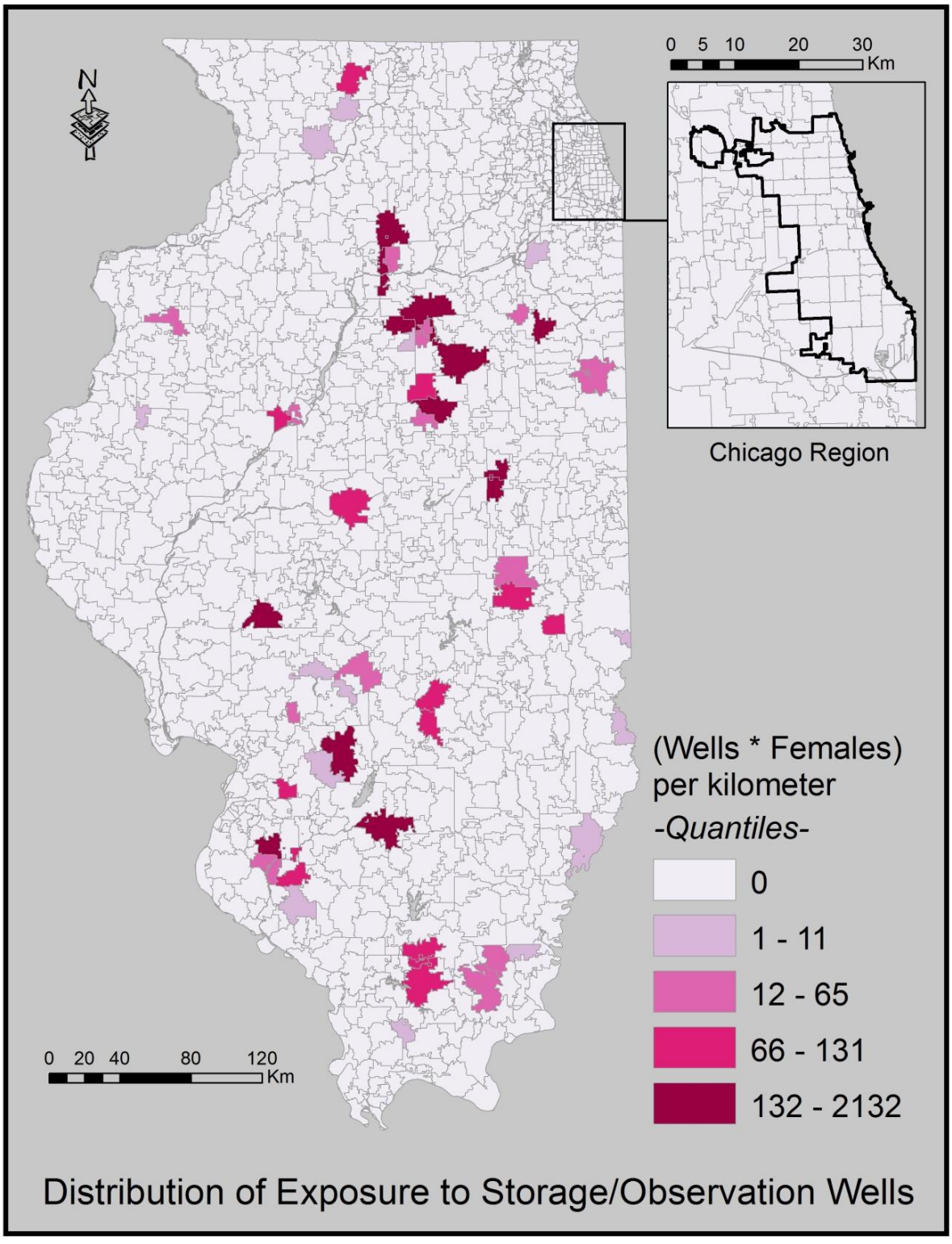

N
0 5 10 20 30
Km
Chicago Region
(Generators * Females)
per kilometer
-Quantiles-
0
1 - 33
34 - 308
309 - 1271
1272 - 17049
0 20 40 80 120
Km
Distribution of Exposure to Hazardous Waste Generators

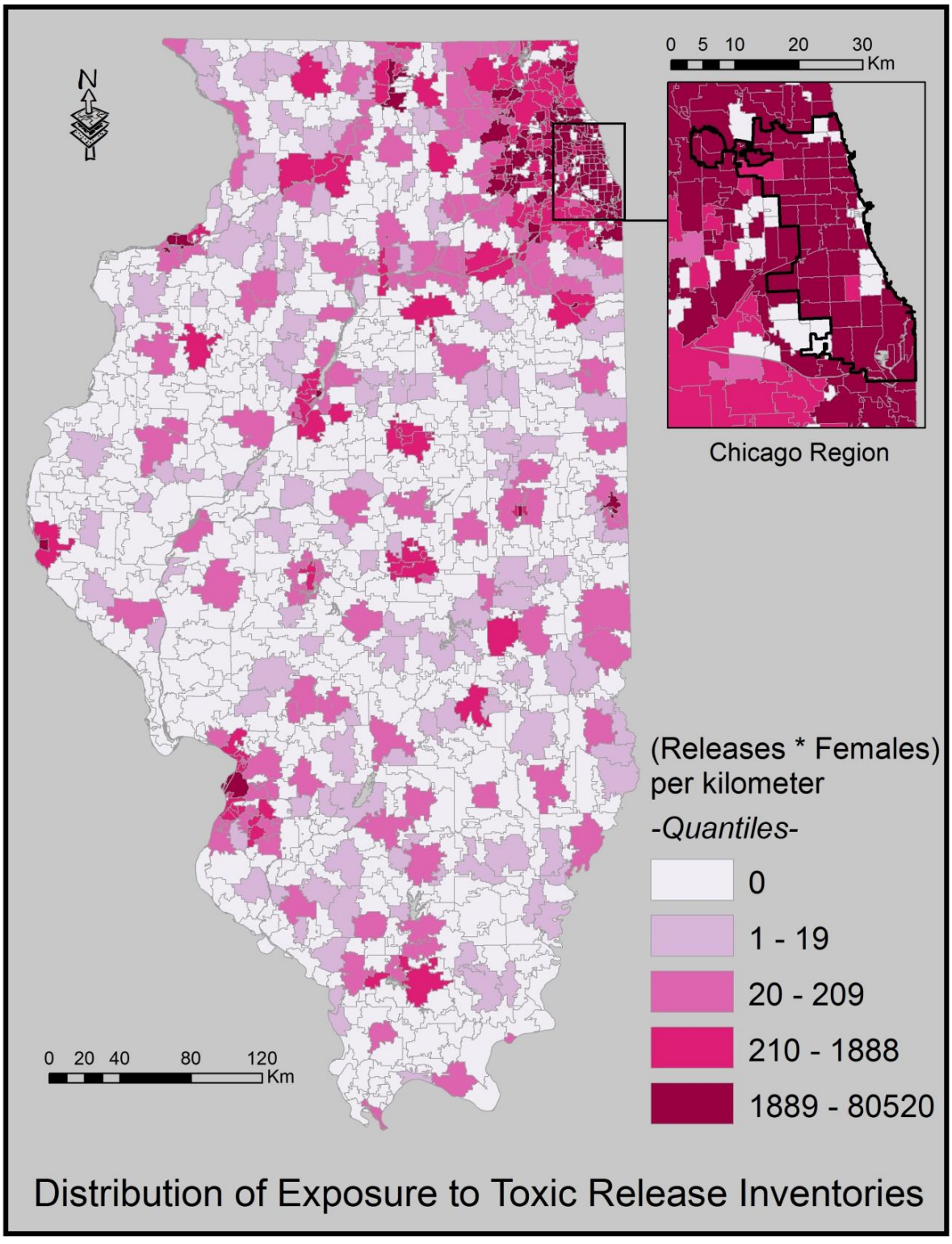

N
0 5 10 20 30
Km
Chicago Region
0 20 40 80 120
Km
(Facilities * Females)
per kilometer
-Quantiles-
0
1 - 4
5 - 9
10 - 54
55 - 9015
Distribution of Exposure to Pesticide Producing Facilities

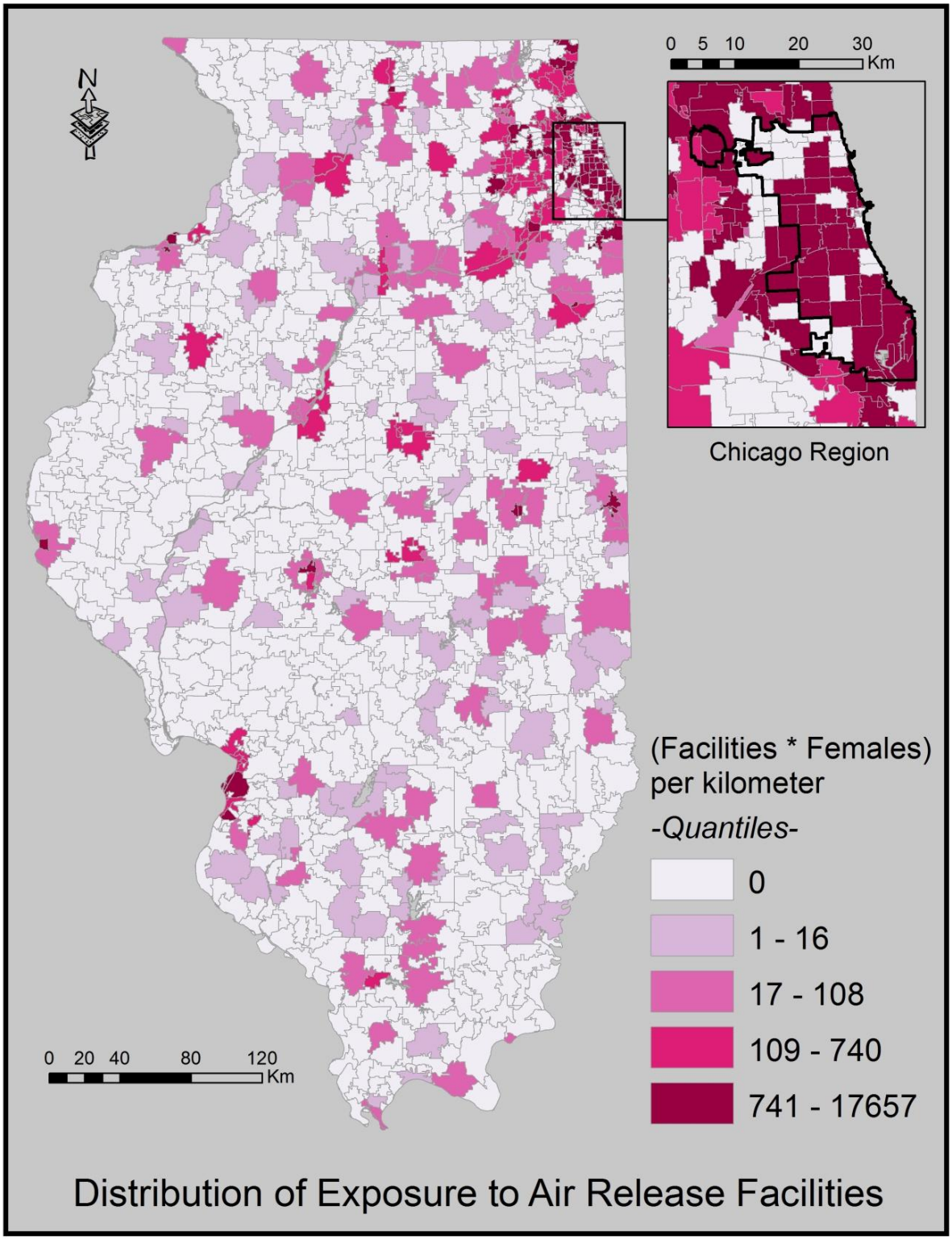

N
0 5 10 20 30
Km
Chicago Region
0 20 40 80 120
Km
Wells per Female
-Quantiles-
0.00 - 0.02
0.03 - 0.10
0.11 - 0.19
0.20 - 0.31
0.32 - 2.01
Distribution of Exposure to Water Wells

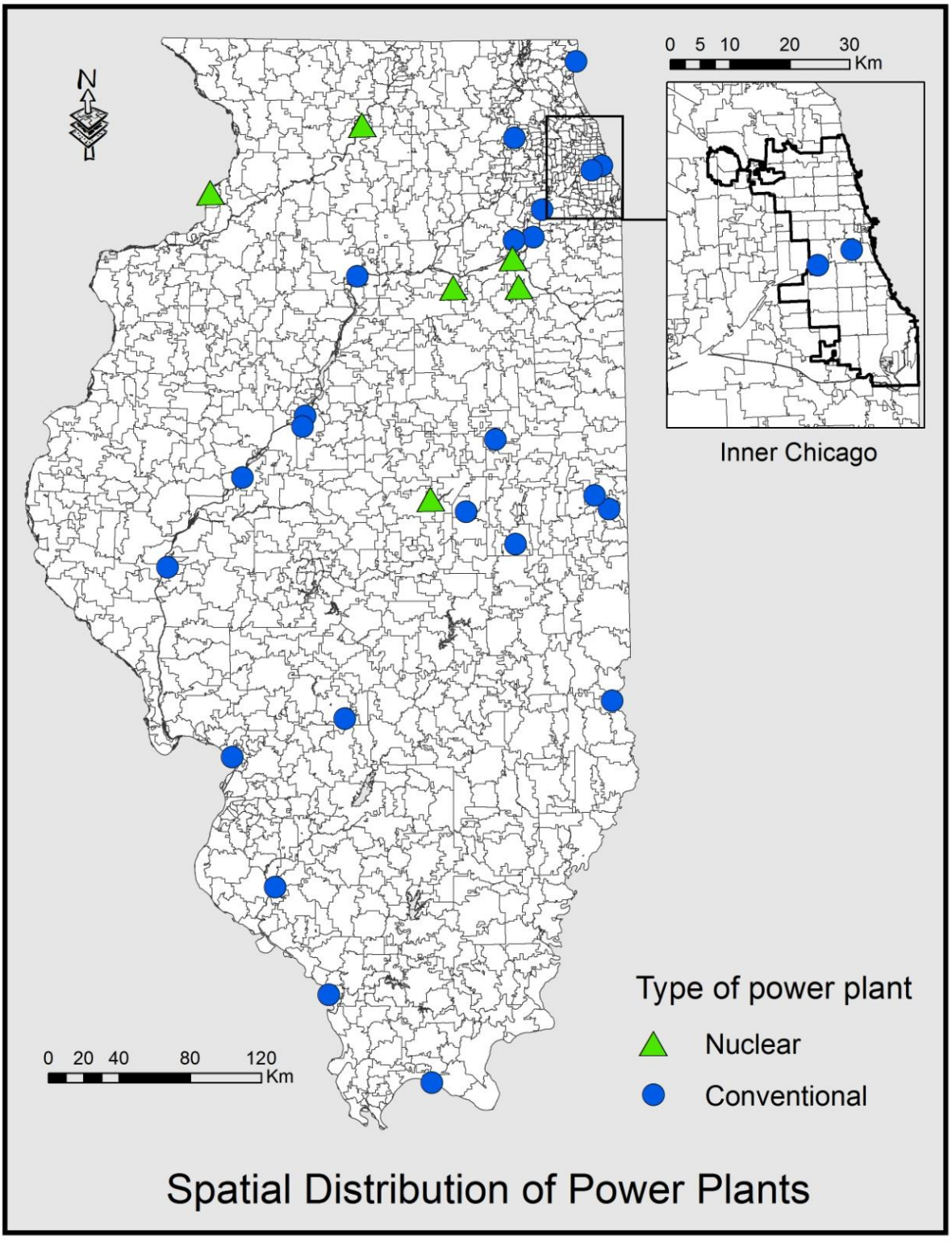

N
0 5 10 20 30
Km
Chicago Region
0 20 40 80 120
Km
(Wells * Females)
per kilometer
-Quantiles-
0
1 - 6
7 - 19
20 - 91
92 - 1443
Distribution of Exposure to Oil/Gas Injection Wells

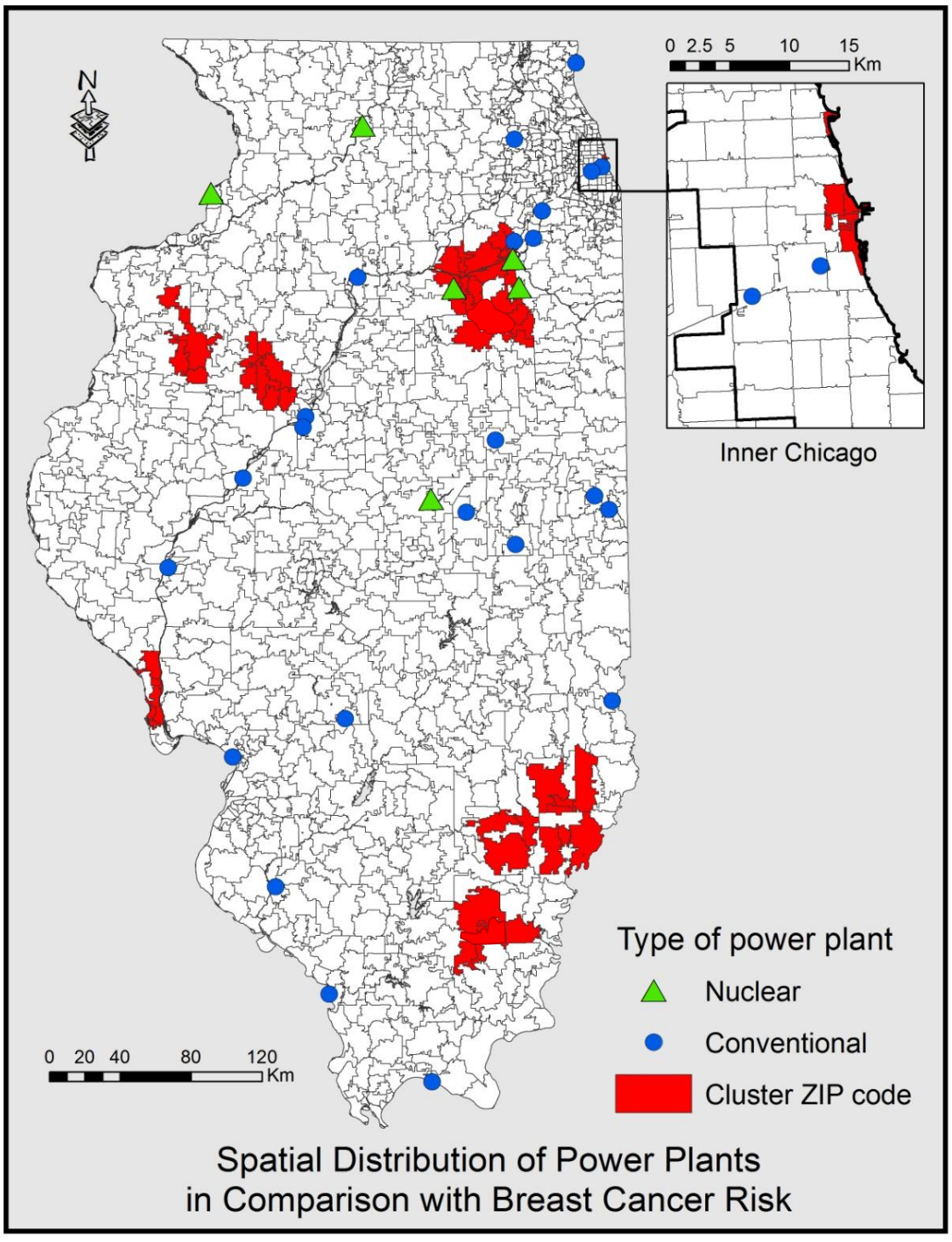

N
0 5 10 20 30
Km
Chicago Region
0 20 40 80 120
Km
(Releases * Females)
per kilometer
-Quantiles-
0
1 - 19
20 - 60
61 - 266
267 - 1609
Distribution of Exposure to Wastewater Releases


N
0 5 10 20 30
Km
Chicago Region
(Wells * Females)
per kilometer
-Quantiles-
0
1 - 11
12 - 65
66 - 131
132 - 2132
0 20 40 80 120
Km
Distribution of Exposure to Storage/Observation Wells

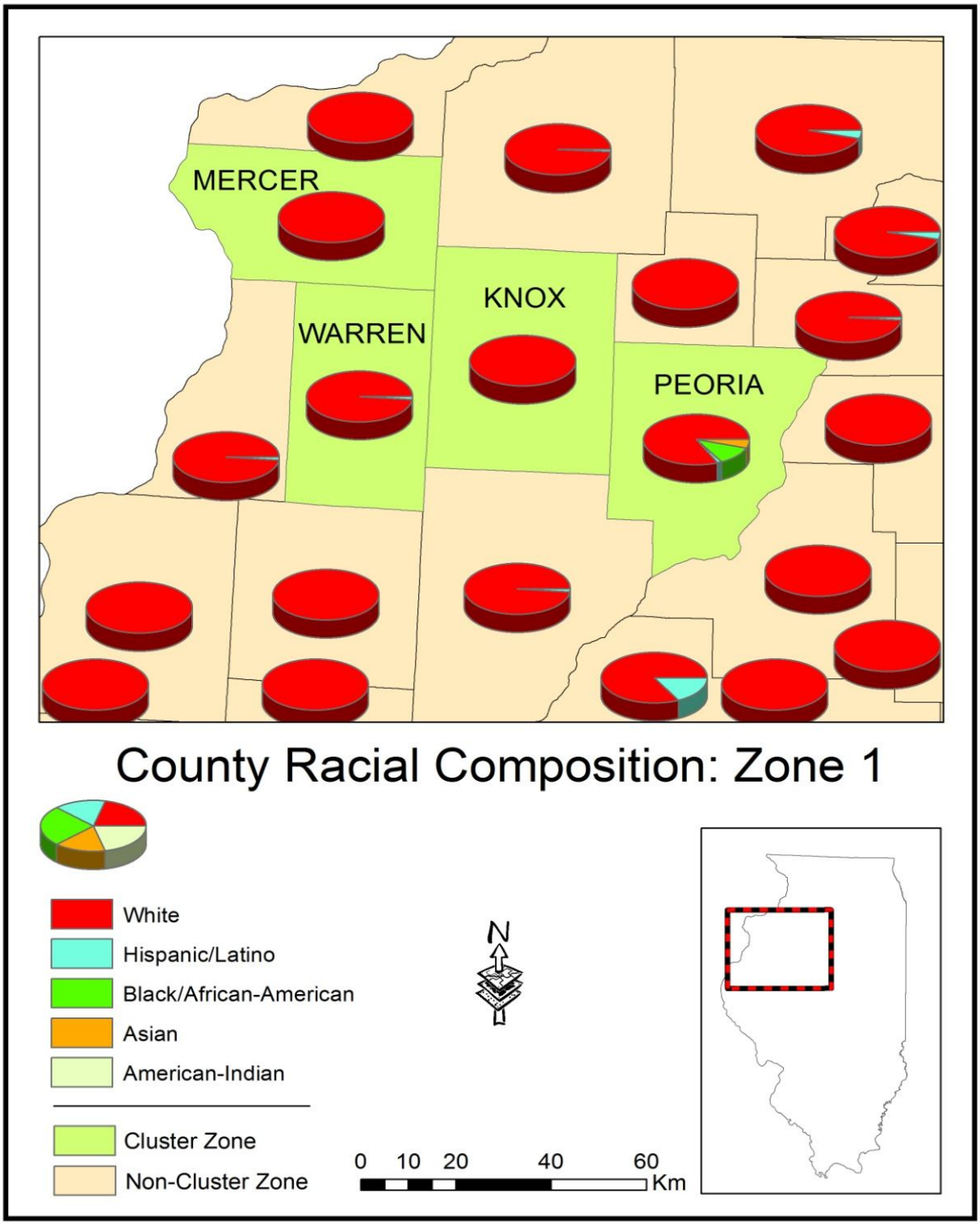

N
0 5 10 20 30
Km
Chicago Region
0 20 40 80 120
Km
(Releases * Females)
per kilometer
-Quantiles-
0
1 - 19
20 - 209
210 - 1888
1889 - 80520
Distribution of Exposure to Toxic Release Inventories

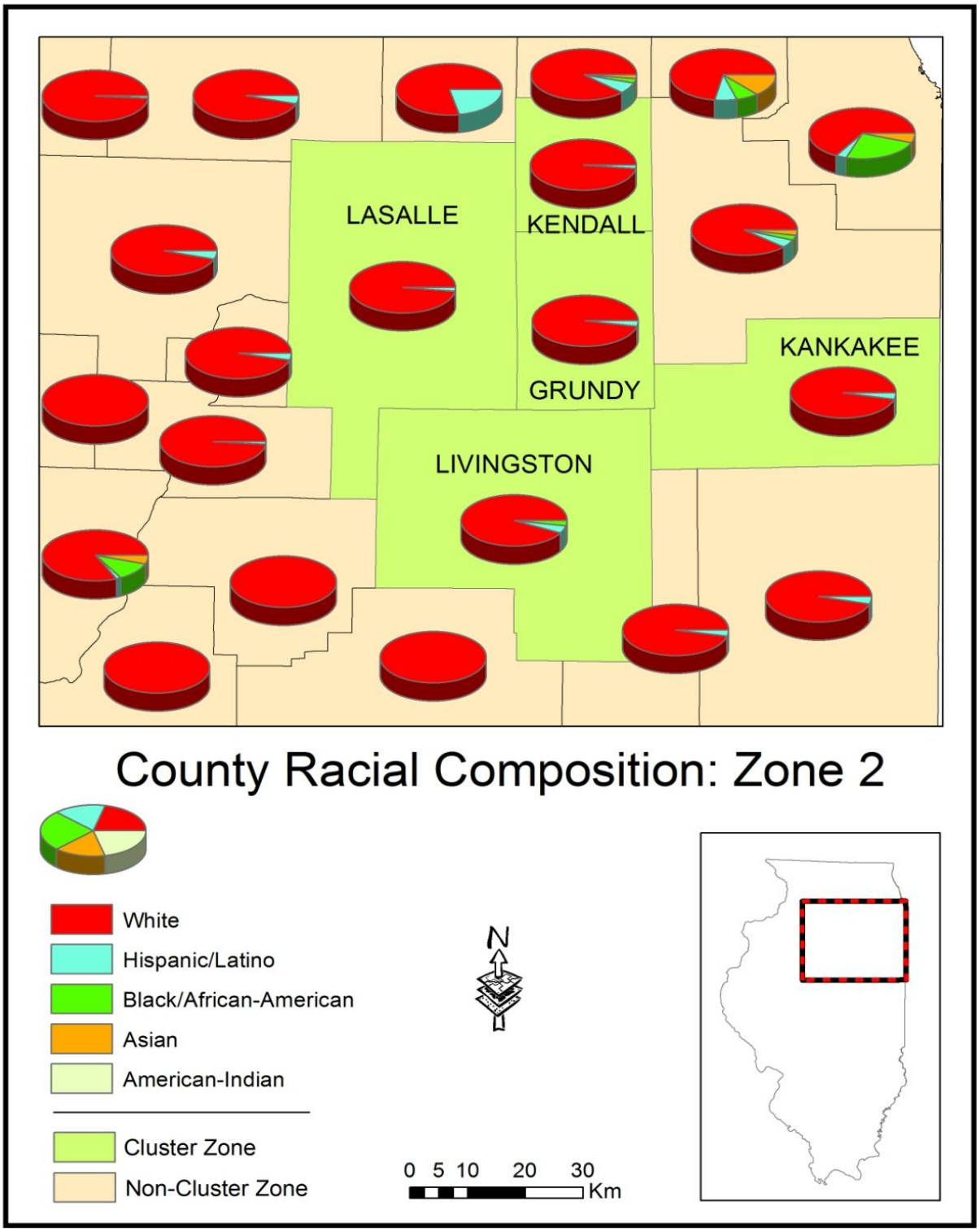

N
0 5 10 20 30
Km
Chicago Region
0 20 40 80 120
Km
(Facilities * Females)
per kilometer
-Quantiles-
0
1 - 16
17 - 108
109 - 740
741 - 17657
Distribution of Exposure to Air Release Facilities

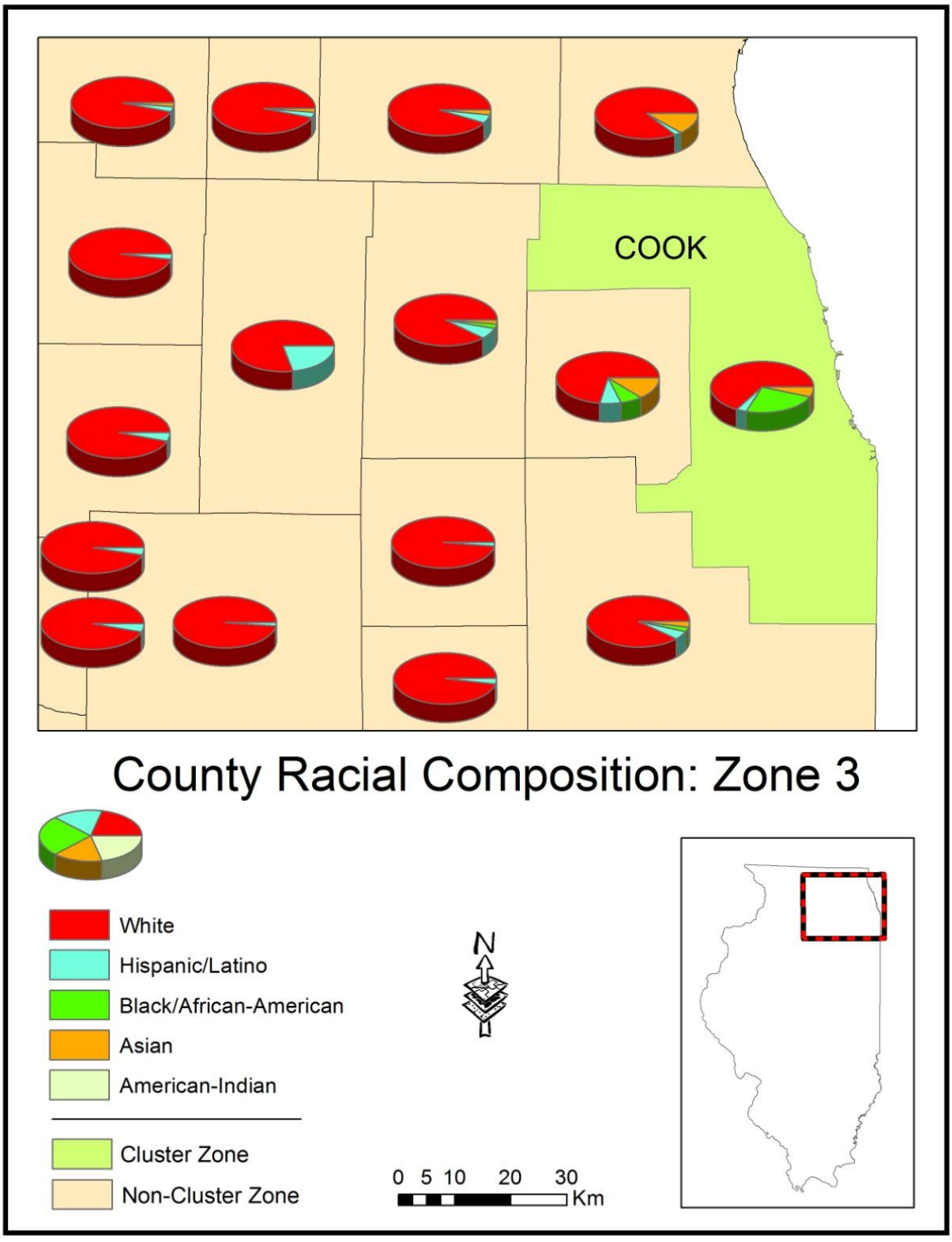

N
0 5 10 20 30
Km
Inner Chicago
0 20 40 80 120
Km
Type of power plant
Nuclear
Conventional
Spatial Distribution of Power Plants

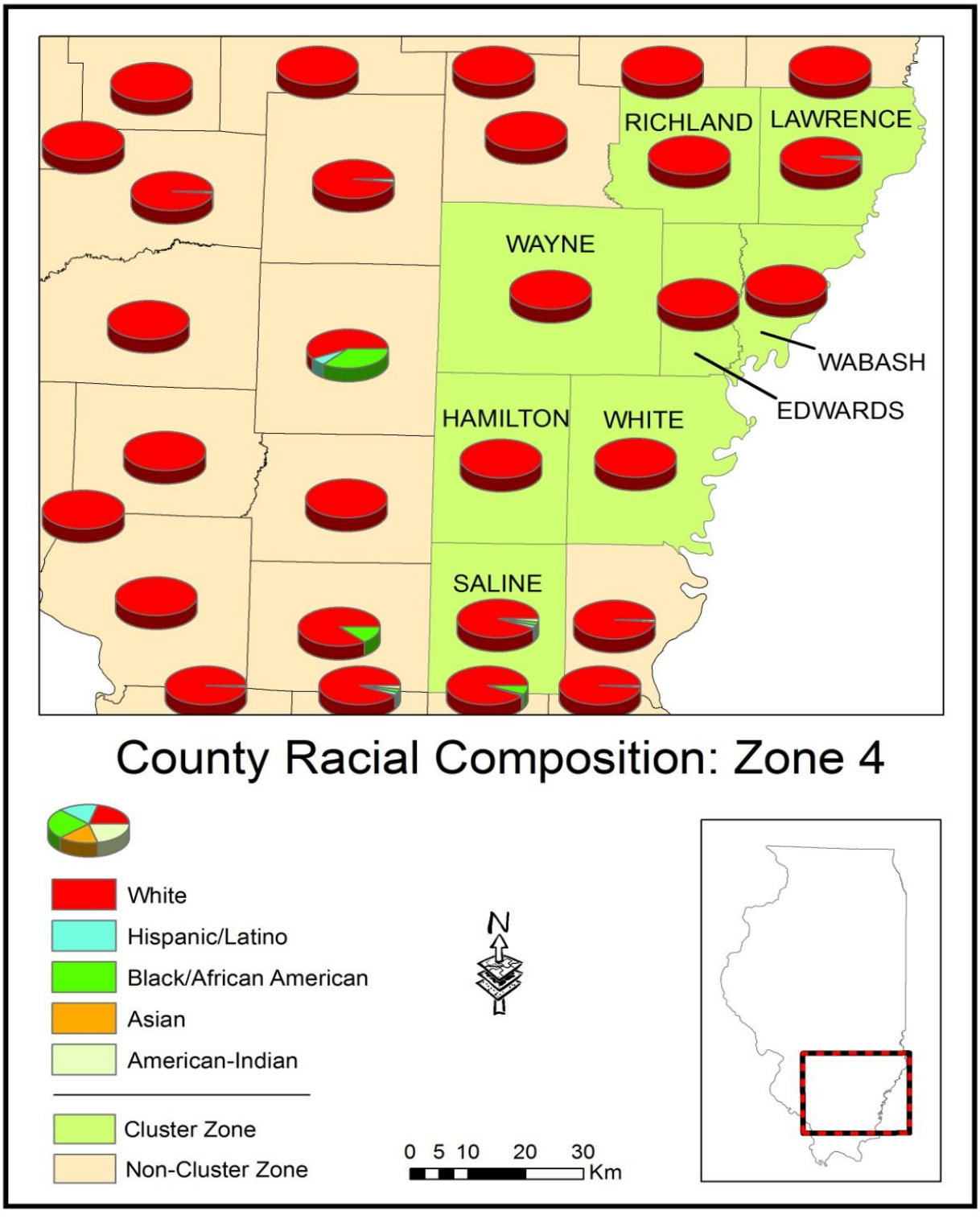

N
0 2.5 5 10 15
Km
Inner Chicago
0 20 40 80 120
Km
Type of power plant
Nuclear
Conventional
Cluster ZIP code
Spatial Distribution of Power Plants
in Comparison with Breast Cancer Risk

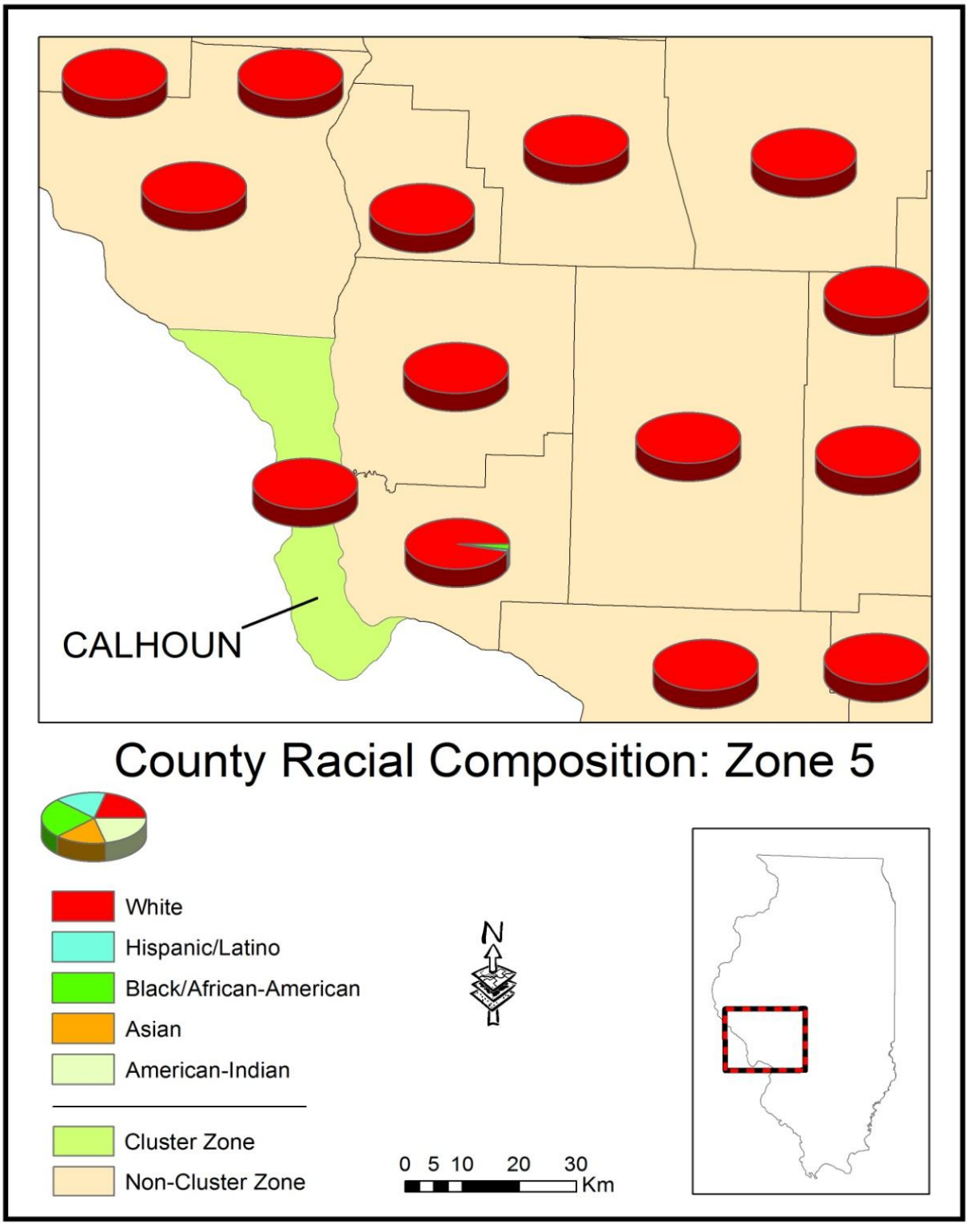
APPENDIX B


MERCER
WARREN
KNOX
PEORIA
County Racial Composition: Zone 1
White
Hispanic/Latino
Black/African-American
Asian
American-Indian
Cluster Zone
Non-Cluster Zone
N
0 10 20 40 60
Km


LASALLE
KENDALL
KANKAKEE
GRUNDY
LIVINGSTON
County Racial Composition: Zone 2
White
Hispanic/Latino
Black/African-American
Asian
American-Indian
Cluster Zone
Non-Cluster Zone
N
0 5 10 20 30
Km


COOK
County Racial Composition: Zone 3
White
Hispanic/Latino
Black/African-American
Asian
American-Indian
Cluster Zone
Non-Cluster Zone
N
0 5 10 20 30
Km


RICHLAND
LAWRENCE
WAYNE
WABASH
EDWARDS
HAMILTON
WHITE
SALINE
County Racial Composition: Zone 4
White
Hispanic/Latino
Black/African American
Asian
American-Indian
Cluster Zone
Non-Cluster Zone
N
0 5 10 20 30
Km


CALHOUN
County Racial Composition: Zone 5
White
Hispanic/Latino
Black/African-American
Asian
American-Indian
Cluster Zone
Non-Cluster Zone
N
0 5 10 20 30
Km