SPATIAL ATTRIBUTES OF BLOOD ORGANOCHLORINE CONCENTRATIONS IN WASHINGTON COUNTY, MARYLAND, 1974-1989

by

Shannon L. Henshaw

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland

March 23, 2004

Abstract

People living near potential sources of persistent organochlorine compounds may be exposed to higher than background levels of these ubiquitous pollutants. Geographic information is commonly used to evaluate these pollutants in the environment, but overlooked in studies modeling these compounds in humans. The overall goal of this thesis is to evaluate residential location as a potential exposure determinant by investigating the geographic distribution of human biomarkers of exposure to persistent organochlorine compounds in a community with a potential source. The community chosen for this research was Washington County, Maryland because it contains a Superfund site contaminated with organochlorines and a large proportion of the residents had donated their blood for research purposes.

To determine if the Superfund site was a potential source of exposure to County residents, 110 soil samples collected in and around the site were used to predict soil levels of 1,1-dichloro-2,2-*bis*(chlorophenyl)ethylene (DDE) at unsampled locations with geostatistical methods. Potential risk factors of exposure to organochlorines were then evaluated using generalized least squares regression to model blood samples of 2,184 residents. Participant addresses were geocoded so potential spatial risk factors such as residential distance to the site could be examined. To evaluate the sensitivity of these models to positional bias in the geocoding process, 163 addresses were visited and the distance between the geocoded coordinates and the residence was determined. Conditional simulation was used to estimate this distance in the remaining geocoded addresses, and the robustness of the regression models was tested using a Monte Carlo approach.

The results provide suggestive evidence that the Superfund site is a potential source of organochlorine exposure to surrounding residents. A statistically significant association between blood dieldrin levels and residential distance from the site adjusting for age, gender, smoking status, education and drinking water source was found, although no other associations between the site and other organochlorines were found. The dieldrin association is robust to geocoding positional bias. Overall, spatial information is found to be important for deriving geographic-based predictors of blood organochlorine levels and accounting for spatially dependent model residuals and should be evaluated when dealing with human exposure and biomarkers.

Thesis Readers

Patrick N. Breysse, Ph.D., Advisor; Department of Environmental Health Sciences

Frank C. Curriero, Ph.D.; Co-advisor; Department of Biostatistics

Gregory E. Glass, Ph.D.; Department of Molecular Microbiology and Immunology

Paul T. Strickland, Ph.D.; Department of Environmental Health Sciences

Kathy Helzlsouer, M.D.; Department of Epidemiology

Jacqueline Agnew, Ph.D.; Department of Environmental Health Sciences

Acknowledgments

This study was funded by the Johns Hopkins National Institute for Environmental Health Sciences (NIEHS) Center in Urban Environmental Health (P30 ES 03819). My work as a student at the Johns Hopkins Bloomberg School of Public Health was supported by the NIEHS training grant (ES 07141) and the National Institute for Occupational Safety and Health Education and Research Center training grant (T42-CCT310419). These funds are greatly appreciated, as this research opportunity would not be possible without them.

I would first and foremost like to thank my co-advisors, Dr. Patrick N. Breysse and Dr. Frank C. Curriero for their endless support and guidance throughout this research. Not only did they offer their knowledge and expertise in this field, but they were excellent mentors, both academically and personally. For this, I am most appreciative. All the skills I have gained through this research are a result of their teaching.

I am indebted to Dr. Frank C. Curriero for his overwhelming patience and commitment to this research. Not only did he spend long hours working with me on the data analysis portion of this research, but he also accompanied me on several trips to Hagerstown to help with data collection. When problems were encountered throughout this project, his motivation kept my spirits high.

In addition, I would like to thank Dr. Paul T. Strickland, Dr. Greg E. Glass, Dr. Kathy Helzlsouer, and Dr. Jacqueline Agnew for serving on my thesis advisory committee. Their suggestions and advice as how to better my research were greatly appreciated.

I would also like to thank the staff at the Johns Hopkins Training Center for Public Health Research in Hagerstown, Maryland for their help in attaining the data necessary for this research. I am gracious for their hospitality as I spent many days working in their office. Furthermore, Sandy Hoffman and Alyce Burke were very efficient in making sure I had the information I needed, and for this, I am also grateful.

In addition, I would like to thank Tim Shields of the Department of Molecular Microbiology and Immunology for imparting some of his vast knowledge of Geographic Information Systems on me, and Humberto Monsalvo, of the United States Environmental Protection Agency for his helpful cooperation in obtaining environmental data documents and guidance along visits to the Superfund site.

I would like to express my sincere gratitude to my colleagues and friends within the School. Their friendship, advice, and encouragement are uplifting everyday, and a constant reminder of what is important in life. I would especially like to thank my fiancé, Brian, for his support throughout this degree program. He remained by my side whether it was out in the field collecting data or working beside me while I wrote. His patience was astounding despite the fact that many evening and weekend hours were spent on this research, and many sacrifices were made on his part.

Finally, I would like to thank my parents for instilling in me the confidence and the drive necessary to succeed. Without their guidance throughout my life, and their support of my education, this degree could not be achieved. I only hope that I can use the skills that I have learned to help and encourage others as they did me.

Table of Contents

List of Tables

List of Figures

INTRODUCTION

1. Introduction

Geographic or spatial information has long been used to study the environmental contamination patterns of pollutants. For instance, levels of contamination are often measured in multiple media surrounding a source in order to determine the environmental fate, transport, and transformation of the pollutants (*1*). Many of these well-studied environmental contaminants can also been detected in humans, indicating exposure to these compounds. Therefore, the importance of spatial information may extend beyond environmental data to measures of human internal dose. There have been few studies characterizing the geographic distribution of human biomarkers of exposure to these substances (*2*).

Persistent organochlorine compounds are examples of well-studied ubiquitous environmental pollutants that are also detected in humans. Although they are found virtually everywhere, research involving the geographic distribution of these chemicals in the environment suggests that residing near a source may be a risk factor for higher than background exposure to organochlorines (*3,4*). Some organochlorines, such as dichlorodiphenyltrichloroethane (DDT), are classified as possible human carcinogens (Group 2B) by the International Agency for Research on Cancer (IARC). Polychlorinated biphenyls (PCBs) have also been linked to several forms of adverse health effects, and have been classified as Group 2A, probable carcinogens by IARC (*5*). These substances are known to bioaccumulate in the adipose tissue of humans and therefore, pose a threat to those exposed (*6*).

Hagerstown, Maryland, located in Washington County, contains the Central Chemical National Priority List (NPL) site. Central Chemical Company (CCC) produced pesticides, such as DDT, and fertilizers from the 1930's until the 1980's. Hence, it is a potential source of organochlorine exposure to Washington County residents. The release of organochlorine compounds from the site has been partially characterized by the Maryland Department of the Environment (MDE) and the United States Environmental Protection Agency (EPA) (*7-10*). Additionally, the human internal dose of these chemicals in residents of Washington County can be evaluated using a large database of blood organochlorine data collected as part of two population-based epidemiological studies (CLUE I and CLUE II) conducted in Washington County.

1.1. Research Purpose

The goal of this research is to evaluate residential location as a potential exposure determinant by investigating the geographic distribution characteristics of human biomarkers of exposure to persistent organochlorine compounds in a community with a potential source. In order to accomplish this goal, the following specific aims were proposed:

1. To gather all organochlorine environmental contamination data collected in Washington County, and determine whether or not the Superfund site in Washington County is a potential source of organochlorine exposure.
2. To characterize the geographic distribution of all the blood organochlorine levels as internal dose measures, and explore spatial patterns in both the 1974 and 1989 CLUE data.
3. To evaluate the influence of established risk factors, such as gender, age, smoking status, body mass index (BMI), dietary intake, occupation, breastfeeding status,

and consumption of well vs. municipal water on blood organochlorine levels, and then further account for possible spatially dependent residual variation using other spatial variables such as urban vs. rural residence, and the distance from the residence to the Superfund site in both the 1974 and 1989 data.

4. To determine the sensitivity of the final statistical models developed in Specific Aim Three to positional inaccuracies in the geocoding process.

This research tests the following hypotheses:

1) The CCC Superfund site is a potential source of organochlorine exposure to surrounding residents.
2) Spatial determinants can improve statistical models of individual blood organochlorine biomarker levels that are known to depend on age, gender, BMI, and other individual characteristics.
3) There will be a spatial relationship between the blood levels of those organochlorines produced/used by the CCC and the Superfund site location, while the spatial patterns of other ubiquitous organochlorine compounds that were not produced by the CCC will not be related to the Superfund site location.
4) The risk factors for, and spatial distribution of human blood organochlorine biomarker levels in Washington County, Maryland will differ in 1974 compared to 1989.
5) Spatial regression models will be robust to positional inaccuracies in the geocoding process.

1.2. Significance

Currently, there are over 1,200 Superfund sites across the country that are contaminated with substances that adversely affect human health (*11*). Furthermore, Superfund sites are often located in urban areas surrounded by residences. Therefore, the health of people living near sites like these is dependant on the ability to accurately characterize and evaluate environmental contamination on and around these sites. Research in assessing the impact of residing near a potential pollution source is lacking. The results of this study help to better characterize residential location as a determinant in organochlorine exposure by determining the relative contribution of spatial information in evaluating human internal dose.

1.3. Organization of Research

This thesis is composed of eight sections. Introduction and background sections are followed by a section describing the general research methods. The specific aims are addressed in sections four, five and six, which contain publishable manuscripts. The subsequent section contains the overall conclusions of the research, followed by a reference section.

The background section contains a literature review on the properties, exposure pathways, health effects and biomarkers of organochlorines as well as a description of spatial statistics, the CLUE campaigns and studies, and a historical overview of the Superfund site. The chapter on research methods briefly describes the research tools used in this study.

The first manuscript entitled, "GIS and Geostatistics: Tools for Characterizing Environmental Contamination" addresses Specific Aim One of the research. In this manuscript, levels of 1,1-dichloro-2,2-*bis*(chlorophenyl) ethylene (DDE) in soil measured throughout the last decade are used to predict levels of DDE in soil at areas where samples were not taken, including areas currently undergoing residential development.

The second manuscript entitled, "The Use of Spatial Information in Determining Potential Risk Factors of Blood Organochlorine Levels in a Population Living Near a Potential Source" addresses Specific Aims Two and Three of the research consisting of regression models accounting for spatial dependence in the data and both spatial and individual risk factors that were determined to be influential in evaluating blood organochlorine levels in either or both 1974 and 1989.

The third manuscript entitled, "Geocoding Positional Bias and its Effect in Analyzing the Spatial Distribution of Blood Dieldrin Levels" describes the methods involved in determining the residential location of study participants and the bias associated with these methods. Furthermore, sensitivity analysis is used to evaluate the effect of these positional biases on the overall research goal, accomplishing Specific Aim Four of the research.

The concluding section summarizes the overall findings of the research, opportunities for future research, and the public health significance of this research. Finally, the reference section contains a bibliography of all cited literature followed by three appendices, a copyright permission letter for the first manuscript, and the curriculum vitae of the author.

BACKGROUND

2. Background

2.1. GIS and Spatial Statistics

Evaluating spatial relationships and geographic determinants of health is a growing area of environmental and public health research. Traditional environmental health studies have evaluated temporal changes in diseases and their determinants. More recently, however, there is a growing interest in evaluating both spatial and temporal patterns of health and their environmental determinants. As a result, geographic information systems (GIS) and spatial statistical techniques have become increasingly popular in environmental health applications and to environmental health practitioners (*12-19*). As GIS becomes more widespread, various applications of GIS have evolved. In its basic form, GIS is a database system with the distinguishing feature that it deals with spatially (or geographically) referenced data. "Where" in addition to "what" that is measured or observed is important and thus recorded and stored in a GIS database.

With location information linked to data values, GIS becomes a visual database providing comprehensive mapping capabilities as it's most basic and fundamental construct. However, the functionality of a GIS extends beyond producing maps. GIS are computer-based systems that serve as powerful tools to input, store, retrieve, manipulate, descriptively analyze, and output data that have geographic characteristics (*2,20*). GIS provides methods for the user to query data, study the relationship between any selected geographical or spatial variables, and generate a comprehensive spatial database. In addition, GIS can serve as a relational database by providing a common format to link specific data to additional geographical or environmental data collected for other purposes in the same geographic area.

In this thesis, organochlorine levels in soil surrounding a potential source are analyzed using GIS and spatial statistics. Furthermore, GIS and spatial statistics are used to evaluate the geographic distribution of spatial determinants and dependence of blood organochlorine levels. These data are spatially referenced because they can be linked to the residential location of the individual from which the blood was drawn. Details about these methods are found in the research methods section.

The area of spatial statistics known as geostatistics is ideally suited for characterizing the geographic distribution of environmental contamination and for evaluating potential spatial determinants. Geostatistics is the set of statistical techniques used to analyze spatially referenced data (*21-23*). Since spatial information, such as place of residence, can be linked to human exposure, geostatistics may also be used to evaluate human internal dose measures of environmental exposure. Originally developed in the mining industry to predict recoverable ore reserves, geostatistics have found widespread use in environmental and public health related applications (*22-26*).

A key concept in geostatistics is that of spatial dependence. Spatial dependence is the condition by which data points that are closer together in space are more similar than those that are further apart. This may apply to internal dose measures of environmental exposure since humans living near one another may have more similar exposures than those living far away from one another. The characterization of spatial dependence is often carried out with a function known as the variogram, which is discussed in more detail in the research methods section.

Specialized software packages exist to perform geostatistical analyses (*27-30*). The open source statistical computing environment R is used in this thesis. R is similar to the S and S-Plus statistical computing package (MathSoft), with main advantages being that it is free and continually expanded with contributed libraries for a wide range of specialized topics. R has an extensive range of built-in statistical techniques that can be used to perform standard statistical analyses from exploratory to more confirmatory model-based analyses. R is compatible on PC, Unix, Linux, and Macintosh platforms. In addition, R has over 300 official open source add-on libraries for other specialized techniques with several of these libraries containing tools to perform geographic/spatial analyses (*28*). Furthermore, R has powerful and unique graphical capabilities as is evidenced in this thesis. R has widespread usage in the academic community and increasing usage in industrial and government sectors. More detailed discussions about R and its contributed libraries are found elsewhere (*31*).

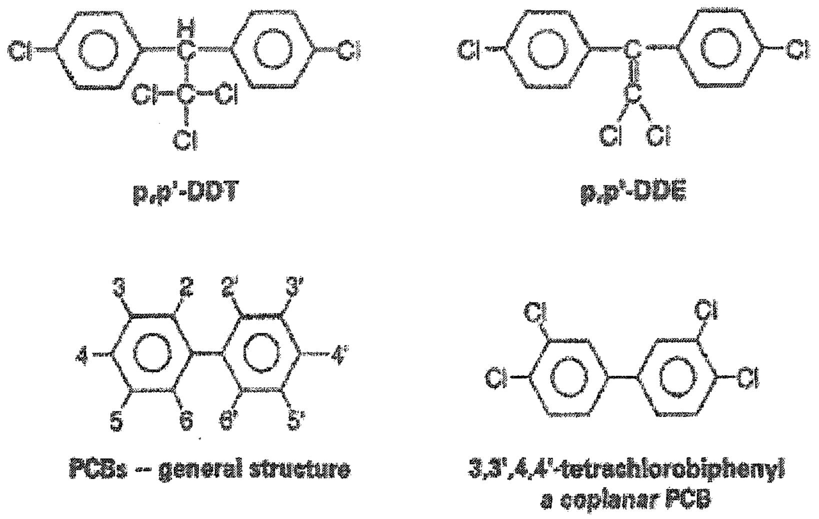
2.2. Organochlorines

Organochlorines are chlorinated aliphatic and aromatic hydrocarbons that are used extensively across the globe. Organochlorines are used as pesticides, insecticides, herbicides, fungicides, preservatives, and industrial fluids (*32-35*). Some organochlorines, such as dioxins, are man-made byproducts of chemical processes. Although many organochlorine residues can be found in humans, and several are detectable in almost everyone, dichlorodiphenyltrichloroethane (DDT) along with its derivatives and polychlorinated biphenyls (PCBs) are thought to be the main contributors to the overall organochlorine body burden in humans (*36,37*). Therefore, although other organochlorine compounds are not ignored in the proposed study, more emphasis will be placed on DDT and PCBs. These compounds are appropriate for this study because DDT represents a chemical that was extensively used at the CCC site, while PCBs represent ubiquitous organochlorines that were not often used at the CCC site.

One of the most persistent organochlorine compounds is DDT. DDT exists as two isomers, p,p'-DDT and o,p'-DDT that differ by the positioning of one chlorine atom (*38,39*). Figure 2.1 contains the chemical structure of DDT. Commercial grade DDT, made up of approximately 85 percent p,p'-isomer and 15 percent o,p'-isomer, was first formulated in 1873, and in 1939, its use as a powerful insecticide was discovered (*38,40*). After World War II, DDT was put to use worldwide. It was applied mainly to cotton, soybean, and peanut crops. In addition, DDT was used to kill the malaria-carrying mosquitoes and typhus-infected lice, and to prevent the spread of yellow fever and sleeping sickness (*32,41,42*). DDT was so effective that the World Health Organization (WHO) gave DDT credit for saving 25 million lives. However, in the 1940's, problems with DDT began to develop. Several insect species had become resistant to DDT, and DDT's toxic impact on some fish species became apparent. In 1972, the United States banned the use of DDT, except for in emergency situations (*43*). Currently, the EPA requires that any spills or releases of more than one pound of DDT to the environment be reported (*6*). DDT is still used in many developing parts of the world and is still the insecticide of choice for killing malaria spreading vectors (*40,42,44*). For example, in Mexico, neighborhoods are sprayed regularly with DDT as part of sanitation campaigns that began in 1960 to control malaria (*45*). The application of DDT to agricultural crops, spraying DDT for malaria control, and environmental cycling are the main sources of DDT releases to the environment.

Another family of persistent organochlorines is PCBs. PCBs are made up of 209 different congeners of chlorobiphenyl (Figure 2.1). They are distinguished by the degree of chlorination. In 1970, over 86 million pounds of PCBs were produced in the United States. They were used as hydraulic fluids, plasticizers, adhesives, fire retardants, dedusting agents, pesticide extenders, inks, lubricants, cutting oils, carbonless reproducing paper, and in heat transfer systems (*33*). By 1977, only 35 million pounds of PCBs were being produced, and the United States Environmental Protection Agency (EPA) banned most uses of PCBs in the United States in 1979 (*46*). Since then, many other countries have also banned its uses (*47*). Currently, the major sources of PCB releases to the environment are environmental cycling of PCBs previously released, PCB-containing landfill leakage, municipal trash and sewage incineration, and improper disposal of PCB materials. Between 1987 and 1993, the EPA estimated that over 74,000 pounds of PCBs were released to the land (99 percent) and water (one percent) (*46*).

Figure 2.1: Chemical structures of DDT, DDE and PCBs, copied from reference (36)



2.2.1. Environmental Pathways

2.2.1.1. Fate

Organochlorines such as DDT and PCBs are man-made, ubiquitous, eco-toxic persistent organic pollutants (POPs). The higher the degree of chlorination in their chemical structure allows them to be more persistent, bioavailable, and mobile than other chemicals. Since these POPs have physical and chemical properties that enable them to persist in the environment, they can be transported long distances across the globe (*1*).

2.2.1.2. Transport

Organochlorines attach to soil particles strongly via adsorption (*6,46*). DDT has a half-life of two to fifteen years in soil (*6*). The higher the degree of chlorination of the

PCB congener, the stronger it will bind to soil. PCBs do not tend to leach from soil into water unless organic solvents are present (*46*).

Organochlorines do not dissolve easily in water (*6,46*). The higher chlorinated PCBs are less water-soluble than the less chlorinated. PCBs released into water will tend to adsorb to sediment and suspended matter, and eventually may re-enter the water column. If PCBs do not have anything to adsorb to in an aqueous environment, they may volatilize (*46*).

Due to their low vapor pressures, persistent organochlorines do not readily volatilize (*48*). Airborne DDT does not persist in air for extended periods of time (*6,46*). DDT has a half-life of two days in the air (*6*). PCBs released to the atmosphere will primarily exist in the vapor phase. In the vapor phase, PCBs transform, and have half-lives ranging from 12.9 days to 1.31 years. The higher chlorinated congeners will tend to adsorb to airborne particles (*46*). Physical removal of DDT and PCBs is accomplished via both wet and dry deposition (*6,46*).

Persistent organochlorines have non-polar chemical structures and large fat-water partition coefficients (*5,48*). Therefore, they are highly lipophilic, and tend to bioaccumulate in the food chain, making them eco-toxic (*6,46,47*). Over 95 percent of the United States population has detectable levels of DDT and PCBs in their blood (*4,36*). If steady state ingestion of DDT continues over time, a build up of DDT in the body will occur (*49*). If ingestion of DDT were to completely cease, it would take approximately seven to eight years for half of the DDT ingested to be excreted from the body (*38,40,49*). PCBs are also deposited and stored in the body for long periods of time. They have half-lives ranging from one to fifteen years in humans, depending on the congener. The chlordane isomers, mirex, aldrin, and dieldrin, have a five-year human half-life, while hexachlorocyclohexane (HCH) has a human half-life of six years (*38*).

2.2.1.3. Transformation

Organochlorine environmental transformation can occur through biotransformation, abiotic oxidation, hydrolysis, and photolysis. The rates of transformation depend on the structure of the chemicals, the environmental conditions, the media in which the chemical exists, and the concentration of the chemical (*1*).

DDT can leave water and soil by evaporation, sunlight, or microbiological degradation. In soil, DDT usually breaks down into two derivate forms, 1,1-dichloro-2,2-*bis*(chlorophenyl) ethylene (DDE), a product of dehydrogenchlorination, and 1,1-dichloro-2,2-*bis*(p-chlorophenyl) ethane (DDD), a product of dechlorination (*5,6*). DDE also exists as both p,p'- and o,p'-isomers (*38*). Although slow, PCBs are able to ultimately degrade via biodegradation in water and soil as well. In air, PCBs in the vapor phase may react with hydroxyl radicals which allow them to persist in the atmosphere (*46*).

2.2.2. Geographical Distribution

Organochlorines, such as DDT and PCBs, are ubiquitous. They are found throughout the world contaminating aquatic and terrestrial environments (*50,51*). They have even been found in pristine areas, without an apparent source. Because of the physical and chemical properties of these organochlorines, they are able to travel long distances from their source, while resisting both chemical and biological degradation (*5*).

The major mechanism for the mobility of persistent organochlorines may be a cyclical process of being deposited onto the Earth's surface from the atmosphere and then returning to the atmosphere via evaporation. In the 1930s through the early 1970s, during peak organochlorine production in North America and Europe, the cycle may not have been balanced. Organochlorines were mostly being deposited from the atmosphere (where they were released) onto the land and water bodies via wet and dry deposition. However, after the 1970s, the process may have reversed due to the many bans and controls placed on organochlorines in North America and Europe. More organochlorines may have evaporated into the atmosphere from the land and water surfaces along with dust and water vapor than been deposited (*5*).

Since it has been over 30 years since the peak production of organochlorines, the cycle may be more balanced. The concentrations of most persistent organic pollutants are becoming relatively uniform throughout most of the globe. The only exceptions to this are: areas near a current source such as in underdeveloped nations that continue to produce and use organochlorines, areas near a significant recent past source such as the Great Lakes and Hudson Bay regions where they continue to bear heavy organochlorine burdens from historical uses, and the Polar Regions which act as sinks for these chemicals due to global wind patterns (*3,42,51-55*). Furthermore, there has been evidence suggesting organochlorine cycling is influenced by seasonal temperature changes (*5,56*).

2.2.3. Exposure Pathways

2.2.3.1. Inhalation

Due to their low volatility, vapor concentrations of organochlorines in air are usually low, and do not pose a health risk (*6,46*). In fact, it is estimated that exposure to DDT in air represents less than one percent of the average total daily intake of DDT (*5*). However, direct exposure to aerosolized organochlorines has been reported to cause fatal arrhythmia of the heart (*57*). Furthermore, air in areas near waste sites, organochlorine factories or landfills that are contaminated with organochlorines may contain relatively higher levels of organochlorines (*3,58*). In addition, inhaling soil particles near these areas may also be a source of exposure (*6,46*).

2.2.3.2. Ingestion

Ingestion may be the most important pathway of organochlorine exposure for individuals not occupationally exposed to organochlorines (*13,36,59,60*). Levels of these compounds are often found in root and leafy vegetables, dairy products, meat, fish and poultry (*59,61-65*). In a study by Fitzgerald *et al.*, average concentrations of total PCBs in fish from the St. Laurence river and its tributaries ranged from 0.17 to 20.55 ppm wet weight (*66*). Food products produced in the United States usually contain relatively low levels of DDT (*65*). Higher DDT levels are more commonly found in food that has been imported from a country which still uses DDT as a pesticide (*6*).

The most significant source of organochlorine exposure for infants may be breast milk. Organochlorines accumulate in adipose tissue, and in an exposed woman, they may accumulate in the fatty breast tissue, and be excreted in the milk (*67-70*) . Furthermore, fetuses can also be exposed to organochlorines via transmission across the placenta (*5,6,68,70-73*).

Levels of organochlorines in drinking water are usually low since they have low solubility in water. However, drinking water near waste sites, landfills, or areas where they are sprayed for pest control may be contaminated with organochlorines (*74*). In addition, swallowing soil particles near these areas may also be a source of exposure (*5,6,46*).

2.2.3.3. Dermal Absorption

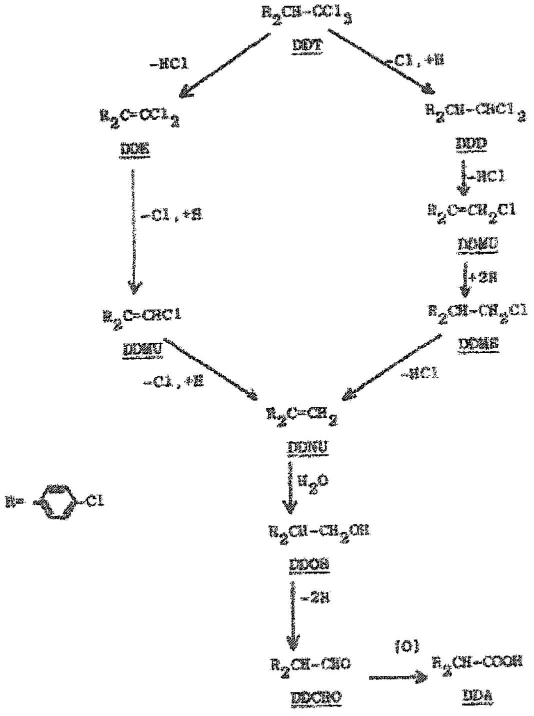
Organochlorines enter the body through the skin by direct contact (*5,75-80*). This may be a significant exposure pathway for those exposed to organochlorines on the job. For example, Lees *et al.* found that transformer maintenance and repair workers came into contact with PCBs quite frequently. They found that contact with PCBs came from not wearing proper personal protective equipment (PPE), handling and not cleaning contaminated tools, not washing hands before eating, and not properly handling, cleaning, and removing contaminated clothes and PPE (*75*). Furthermore, dermal absorption may also be an influential route of exposure for the families of exposed workers, active gardeners, and individuals that swim in natural waters (*13,75*).

2.2.4. Metabolic Pathway

2.2.4.1. DDT, DDE, DDD

Because of its complex chemical structure, several specific metabolic pathways for DDT in different organisms have been postulated (*49,81*). However, it is known that inside the body, DDT metabolizes very slowly by breaking down to form DDE and DDD (*82,83*). The human metabolite of DDE, hydroxylated DDE, and *bis*(4-chlorophenyl)acetic acid (DDA), the metabolite of DDD can be excreted in the urine (*5,83,84*). The metabolic pathway of DDT may resemble that in Figure 2.2 from a study by Gold *et al.* (*83*). DDT is eliminated from the body at a rate of approximately one percent of the stored quantity per day. Fasting may cause adipose tissue to mobilize, and thereby, increase the elimination rate (*48,85,86*).

Figure 2.2: Metabolic pathway of DDT, copied from reference (83).



2.2.4.2. PCBs

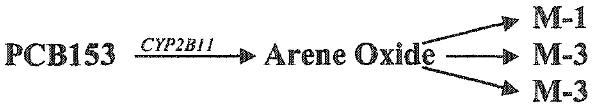
Different congeners of PCBs are metabolized differently. The more persistent PCBs may tend to resist metabolism in many animals and accumulate in fatty tissue. The less persistent PCBs are metabolized via oxidation with the help of cytochrome P-450 enzymes as catalysts. The first step of this process produces the intermediate, arene oxide. Arene oxide is unstable, so it is broken up quickly, and is replaced by a hydroxy group (*87*). The hydroxy group causes the hydroxy derivatives, or metabolites, to be relatively soluble in water, and therefore, enables them to be excreted from the body (*5*). A typical PCB metabolic pathway is as follows:

$$\textbf{PCB congener} \xrightarrow{\textit{P-450-dependent mono-oxygenases}} \textbf{Arene Oxide} \longrightarrow \textbf{Hydroxy metabolites}$$

The metabolism of the less persistent PCB congeners may be due to the fact that they have vicinal H-atoms in the meta-para positions. This may be an important requirement for the formation of arene oxides. Borlakoglu *et al.* found that PCB congeners that have vicinal H-atoms, but not in the meta-para positions, tend to bioaccumulate in fish-eating seabirds, human adipose tissue, and human breast milk (*87*).

The metabolism of one of the most prevalent and persistent PCB congeners found in the blood and adipose tissue of people exposed to commercial PCB mixtures was studied by Ariyoshi *et al* (*37,88,89*). In their study, they found that 2,4,5,2',4',5'-hexachlorobiphenyl (PCB 153) may be metabolized by the cytochrome P-450 isoform, 2B6 (CYP 2B6), in the liver. The major metabolite formed by this pathway in humans was found to be 3-hydroxy-2,4,5,2',4',5'-hexachlorobiphenyl (M-3). Two minor metabolites were also discovered with gas chromatography and electron capture detection, but their structures were not identified. However, there is only a relatively small amount of CYP 2B6 in the human liver, which may be why persistent PCBs are not readily metabolized in humans (*88*).

In dogs and guinea pigs, the metabolism of PCB 153 was better characterized by Ariyoshi *et al.* The metabolites formed via the metabolism of PCB 153 and cytochrome P-450 2B isomers were: 2-hydroxy-4,5,2',4',5'-pentachlorobiphenyl (M-1), 2-hydroxy-3,4,5,2',4',5'-hexachlorobiphenyl (M-2), and M-3 (*88*). The metabolic pathway of PCB153 in dogs may be similar to:



2.2.4.3. Other Organochlorines

Minh *et al.* found that many organochlorines were present in bile indicating that the liver may play a major role in organochlorine excretion (*90*). Organochlorines such as hexachlorocyclohexane (HCH) are biotransformed via dehydrochlorination, glutathione conjugation and aromatic ring hydroxylation. Phenolic products are produced, and excreted (*91*). Aldrin and heptachlor are metabolized to dieldrin and heptachlor epoxide (HE), respectively via oxidation (*48*). Chlordane is metabolized to oxychlordane, while *trans*-nonachlor is hardly metabolized in mammals (*63*). In a fairly recent study, Mehmood *et al.* suggest that hexachlorobenzene (HCB), a persistent hepatic carcinogen, and a commonly used fungicide, may be biotransformed by the human cytochrome P-450 3A4 into pentachlorophenol, and then into tetrachlorohydroquinone in the presence of NADPH (*92*).

2.2.5. Health Effects

2.2.5.1. Acute

DDT is an effective insecticide due to its ability to inhibit neuronal repolarization. This mechanism is also what causes human DDT poisoning (*36*). In humans, ingesting DDT mostly affects the nervous system. Acute effects of large doses of DDT include perioral and lingual paresthesia, apprehension, hypersensitivity, irritability, excitability, dizziness, vertigo, tremors, and seizures (*6,36,93*). Although these symptoms usually do not continue in the absence of continuous exposure, a study by Eriksson *et al.* found that mice exposed to DDT as neonates may have a permanent hyperactive condition as adults (*6,93*). In addition, some studies have shown that short term exposure to DDT may have adverse affects on the reproductive systems. (*5,6,36,67,94,95*). A study by Jonsson *et al.* found that exposure to high levels of DDT and PCBs ceases reproduction in rats (*94*).

Acute exposure to PCBs have been reported to cause acne-like eruptions (chlorine), hyperpigmentation of the skin, spasms, and hearing and vision problems in humans (*36,46,96*). In animals, PCBs have low to moderate acute toxicity. Effects on the liver, kidney and central nervous system (CNS) have been reported (*5,97*).

2.2.5.2. Chronic

DDT is an endocrine disrupter (*98*). Therefore, long term exposure to DDT may affect development and the reproductive system (*94,95,99*). However, the supporting evidence is equivocal. DDT has been suggested to influence premature birth and even spontaneous abortion (*100*). In addition, DDT has been associated with shorter lactation, follicular cysts, and a younger age of natural menopause in females, and reducing bone density, antagonizing the effect of androgen, and reducing sperm concentration, morphology, and motility in males (*36,94,101-103*). Furthermore, long term exposure to DDT may affect liver and thyroid function (*6,13,86,99,104,105*). Occupational studies suggest that long term exposure to DDT may cause reversible changes to enzyme levels in the liver (*6*).

In addition to the many symptoms associated with acute exposure to PCBs mentioned above, long-term exposures to PCBs may cause irritation to the respiratory tract in the form of coughing and tightening of the chest (*33*). PCBs may also be associated with adverse effects in the gastrointestinal tract such as anorexia, weight loss, nausea, vomiting, and abdominal pain (*33*). Both human and animal studies conducted in a variety of species have shown that low background levels of PCBs may adversely affect the immune, cardiovascular, reproductive, nervous, and endocrine systems (*5,36,86,94,104,106-109*). Although the clinical significance is not clear, changes in liver function have also been reported (*5,13,36,46,76*). Specifically, elevated levels of γ-glutamyl transferase (GGT) have been found in the serum of occupationally exposed individuals. Furthermore, a few studies suggest that PCBs induce cytochrome P-450 metabolizing enzymes (*36,97*).

Since PCBs are also endocrine disrupters, they may be responsible for several reproductive and developmental adverse health effects as well. There has been suggestive evidence that occupational exposure to PCBs may cause low birth weights and shorter gestational age (*33*) . In addition, several studies have reported that children of exposed workers have been found to have hypotonia, hyporeflexia and slower motor and cognitive development (*36*). However a study by Lebel *et al.* found no association between organochlorines and endometriosis in women (*110*).

2.2.5.3. Cancer

Most organochlorines are carcinogenic in animals (*38*). The EPA considers DDT, DDE, and DDD to be Group B2, probable human carcinogens (*6*). The International Agency for Research on Cancer (IARC) classifies DDT, DDE and DDD as Group 2B, possibly carcinogenic in humans (*5*). In humans, DDT and its metabolites have been linked with pancreatic cancer, non-Hodgkin's lymphoma, endometrial, and breast cancer (*36,60,69,111-131*). However, the data are not conclusive (*132-136*). In addition, Cocco *et al.* reported increases in liver/ biliary cancer and multiple myeloma prevalence in workers exposed to DDT (*41*). In animals, liver cancer has been found to be associated with DDT ingestion (*5,6,137*). In addition, some studies report lymphomas, lung carcinomas, follicular cell carcinomas, and adenomas of the thyroid as potentially being caused by long term exposure to DDT and its derivatives as well as other organochlorine compounds such as HCB (*5,138*).

Most PCB experts agree that PCBs are probable human carcinogens. IARC classifies PCBs as Group 2A, probably carcinogenic to humans (*5*). The National Toxicology Program (NTP) and the National Institute for Occupational Safety and Health (NIOSH) have concluded that PCBs are probable carcinogens (*106*). Furthermore, after reviewing extensive evidence both from animal and human studies, the EPA considers PCBs as Group B2, probable human carcinogens (*5,106*). Animal studies have shown conclusive evidence that PCBs cause cancer, particularly in the liver (*106*). Occupational and epidemiological studies, although inconclusive, have shown significant relationships between PCB exposure and skin and kidney cancers, and slight associations between cancers in the rectum, liver/biliary, pancreas, prostate, brain, gastrointestinal tract, lung, breast, and lymphoma in those exposed to PCBs (*36,78,118,125,129,139*).

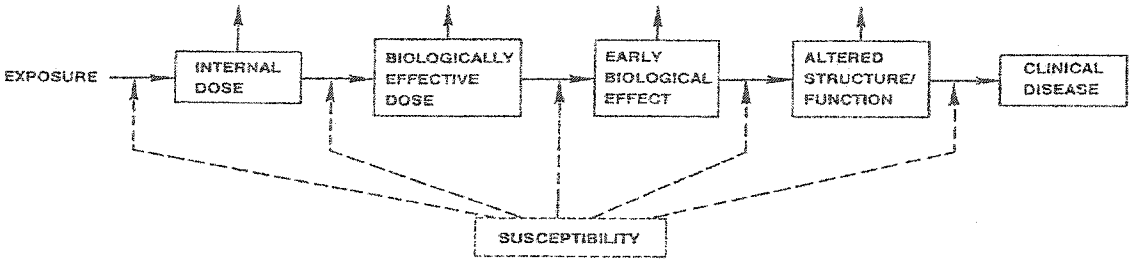
2.2.6. Exposure Guidelines

The Food and Drug Administration (FDA) has set food-specific limits on the amount of specific organochlorine compounds that are allowed in most foods. For example, the FDA's "action level" for DDT is less than five parts per million (ppm) in the edible portion of fish (*140*). In addition, in an attempt to control occupational exposures, the Occupational Safety and Health Administration (OSHA) set permissible exposure limits (PELs) for certain organochlorines. For instance, an employee must not be exposed to an airborne concentration greater than one milligram of DDT per cubic meter of air (1 mg/m^3) in an 8-hour workday within a 40-hour workweek. This PEL is also equal to the American Conference of Governmental Industrial Hygienists' (ACGIH) recommended Threshold Limit Value (TLV) for DDT. Similarly, NIOSH recommends that workers be exposed to less than a half milligram of DDT per cubic meter of air (0.5 mg/m^3), and less than one microgram of PCBs per cubic meter of air (1µg/m^3) during a 10-hour workday in a 40-hour workweek (*6,75*). Furthermore, the EPA set a maximum contamination level (MCL) for PCBs in drinking water at 0.0005 mg per liter of water (*46*).

2.2.7. Biomarkers for Organochlorines

Biomarkers can provide a more accurate measure of exposure than traditional exposure measurements. The Committee on Biological Markers appointed by the National Research Council (NRC) defines biomarkers as indicators of the changes of events in the biological systems or samples throughout the pathway from exposure to disease (*141*). The pathway from exposure to disease, as defined by the NRC, is outlined in Figure 2.3 below. Biomarkers of internal dose reflect how much the body actually absorbs, or the dose the body receives. There are three types of media that are commonly used to measure biomarkers of internal dose from organochlorine exposure. They are: blood, adipose tissue, and breast milk. All three of these media provide long term, cumulative exposure estimates, and are well correlated when adjusted for lipids (*71,129,142-144*). Therefore, organochlorine levels in blood are in steady state with the fatty tissue in which they accumulate. Furthermore, collecting blood samples is much less invasive than collecting adipose tissue or breast milk (*145*). Other biomarkers, such as CYP1A1-dependent enzyme activity and DNA adducts in placenta tissue have also been studied to look at environmental exposure to organochlorines (*53,85*).

Figure 2.3: Toxicological pathway from exposure to disease, copied from reference (141)



2.2.7.1. Analytical Methods

Organochlorines in blood are most commonly analyzed using gas chromatography. Specific congener or isomer detection can be achieved via electron capture detection or mass spectrometry (*38*). In electron capture detection, specific congeners and isomers can be identified by comparing the observed elution times to known standards. In mass spectrometry, specific congeners/isomers are identified by their ion mass to charge ratios, which are characteristic of each compound (*146*).

In the literature, at least two methods of blood serum/plasma extraction have been utilized with gas chromatography. One method is a liquid-liquid extraction that involves extensive cleanup using relatively large quantities of benzene, a carcinogenic solvent. Burse *et al.* reported that this method involved a total of 47 steps, making it extremely labor intensive and time consuming. Furthermore, this method requires the use of custom-packed columns (*147*). This method was used to analyze plasma samples for organochlorines in a study by Helzlsouer *et al.* to determine breast cancer risks (*60*).

The second, and more recommended, method of extracting serum/plasma for gas chromatography is solid phase extraction. Brock *et al.* reported that this method is much faster than the liquid-liquid extraction method, uses commercially available prepacked columns, and uses a smaller quantity of a less hazardous solvent than the liquid/liquid extraction method (*145*). For PCBs, the detection limits of this method range from 0.17 parts per billion (ppb) to 0.45 ppb depending on the congener. The detection limits for the different isomers of DDT and DDE range from 0.54 ppb to 0.66 ppb. Brock *et al.* have reported that there may be interference in the detection of organochlorines in serum due to phthalates also present in serum. This is a limitation in the specificity of the method that may compromise the accuracy of quantification. However, Brock *et al.* reported that the levels of phthalates required to produce such interference are well above those reported to be in average human serum (*145*). Rothman *et al.* and Cantor *et al.* used this method to study the relationship between serum organochlorines and non-Hodgkin lymphoma (*131,148*).

2.2.7.2. Application

Biomarkers of organochlorine exposure have been used in occupational, environmental and epidemiological studies to investigate the risk factors and adverse health effects associated with organochlorine exposure. A summary of a sample of these studies, their methods, and their findings are listed in Table 2.1.

Table 2.1: Summary of the Applications of Organochlorine Biomarkers in Human Studies

Study	Study Subjects	Biomarker	Method of Detection	Study Methods	Study Findings
Burns *et al. 1974* (*149*)	221 human adipose tissues samples obtained during elective surgery at one hospital from 1969-1972	Human adipose tissue analyzed for DDT, DDE, dieldrin, heptachlor epoxide, β-BHC, PCBs	gas-liquid chromatography with electron capture detection	Results were reported, and compared between sexes and race.	DDE levels increased significantly over the time period. These were the highest levels of DDE yet reported in a general population (mean of 17.37 ppm), and second highest DDT (mean of 23.18ppm) reported to date. They did not find a difference in levels of organochlorines between the sexes, but did find that Mexican Americans had significantly higher levels of DDT, DDE, and dieldrin than Anglo-Americans.
Maroni *et al.* 1981(*76*)	80 electrical workers	Blood PCBs (trichloro-biphenyl, pentachloro-biphenyl & total)	Gas chromatography	Personal history; Physical exam; Lab tests	Blood PCB range: 41-1319 μg/kg; Statistically significant positive association between abnormal liver findings & blood PCBs; No association found between chloracne & blood PCBs
Fischbein *et al.* 1982 (*77*)	326 electrical capacitor manufacturing workers	Low (Arochlor 1248) & high (Arochlor 1254) homologs of PCBs in blood (plasma)	Gas chromatography with electron capture detection	Questionnaire; Physical exam; Lab tests; Exposure assessment; t-tests; & non-parametric tests	High prevalence (37%) of dermatological abnormalities was observed; No statistical significant difference between low & high PCB homologs within those subjects with dermatological symptoms
Acquavella *et al.* 1986 (*79*)	205 Employees of an electrical capacitor manufacturing plant	Total PCBs in blood (serum)	Gas chromatography using a modified FDA procedure	Questionnaire; Physical exam; Laboratory analysis; Exposure assessment; Multiple linear regression; correlation analysis	No adverse health effects found; Mean serum PCB level: 18.2 ppb (SD 2.88); Employment duration, cumulative occupational exposure, cumulative fish consumption, & cholesterol level were found to be significant predictors of serum PCB levels
Brandt-Rauf *et al.* 1988 (*150*)	16 municipal workers who clean oil from transformers	PCBs (hepta-, hexa- penta- & tetrachloro-biphenyls & total) in blood (serum)	Reported as "routine techniques"	Personal history; Physical exam; Lab tests	No associations were found with serum PCBs and liver enzymes or serum triglycerides. 6 smokers had abnormalities in the *fes* oncogene related proteins.

Table 2.1: Summary of the Applications of Organochlorine Biomarkers in Human Studies

Study	Study Subjects	Biomarker	Method of Detection	Study Methods	Study Findings
Stehr-Green *et al.* 1988 (*105*)	45 dairy farm family members (from 13 farms) who had consumed heptachlor contaminated raw milk & 94 controls from same region & results from NHANES II as reference population	DDT, DDE, Heptachlor epoxide, oxychlordane, & 6 other organochlorine pesticides in blood (serum)	Methods not specified	Questionnaire (consumption history); Laboratory analyses; Liver function tests; Logistic regression; Chi-squared & parametric statistics	Statistical differences in HE & oxychlordane were found between exposed group & controls, & exposed & reference populations after controlling for age & sex. No difference was found between levels of HE & oxychlordane in control & reference populations. Levels of DDT & DDE were found to be lower in the exposed group. No difference was found in the levels of the other 6 pesticides between the exposed population & the control & reference populations. No change in liver function enzymes was observed.
Stehr-Green *et al.* 1989 (*80*)	5994 people from the NHANES II population	16 organochlorine pesticides in blood (serum)	Gas chromatography with electron capture detection & mass spectrometry	Lab analyses; Descriptive statistics; Bivariate analyses; $\chi 2$ & parametric statistics; Logistic & linear regression	Median DDE levels were found to be 12.6 ppb, & 50% of the samples had greater than 1 ppb DDE. Increasing age, living on a farm, being male may be associated with increased risks of exposure controlling for demographic & seasonal variables.
Angerer *et al.* 1992 (*151*)	53 workers of a municipal waste incinerator and 431 controls	PCBs and HCB in blood (plasma)	Gas chromatography with electron capture detection	Lab analyses; linear correlations as well as non-parametric *U*-tests were used for statistical analyses	Significantly higher levels of HCB were found in workers as opposed to controls, but there wasn't a significant difference in plasma levels of PCBs between workers and controls.
Rivero-Rodriquez *et al.* 1997 (*45*)	40 malaria control workers in Veracruz, Mexico	Isomer specific and total DDT and DDE in abdominal adipose tissue	Gas chromatography with electron capture detection	Questionnaire (work practices & occupational history; Exposure estimation (job tasks); Minor surgery; Linear regression; t-tests;	The geometric mean level of total DDT was found to be 104.48 ug/g. There were statistically significant associations found between higher exposure (based on job task), higher levels of DDT in fat, the use of personal protective equipment, recent weight loss and lower levels of DDT in adipose tissue.

Table 2.1: Summary of the Applications of Organochlorine Biomarkers in Human Studies

Study	Study Subjects	Biomarker	Method of Detection	Study Methods	Study Findings

Rothman *et al.* 1997 (*131*)	74 CLUE participants that have been diagnosed with non-Hodgkin lymphoma (NHL) & 147 matched controls without NHL	DDT, DDE, & PCBs (total and 28 congeners) in blood (serum)	Serum: solid phase extraction with gas chromatography & electron capture detection;	Questionnaire;
				Lab assays;
				Wilcoxon signed rank testing;
				Conditional Logistic regression
					A dose response relationship was found between lipid corrected serum PCB levels and risk of NHL. Seropositivity for the Epstein-Barr virus early antigen may augment the effects of PCBs. No association was found between lipid corrected blood concentrations of DDT and NHL. Median lipid corrected levels of total DDT were found to be 3150 ppb in the cases and 2770 ppb in the controls. Median levels of PCBs were found to be 951 ppb in the cases and 864 ng/g in controls.

Dewailly *et al.* 2000 (*54*)	98 breast-fed and 73 bottle fed Inuit infants from Arctic Quebec, Canada	Breast milk (fat extracts) and blood (serum) analyzed for PCBs, DDE, mirex, HE, chlordane, HCB, endrin, dieldrin	High resolution gas chromatography & electron capture detection	Interviews; clinic visits; lab analyses; t-tests, chi squared tests, and relative risks to evaluate statistical difference between breast and bottle fed infants; logistic regression was used to control for confounders.	The study found that risk of otitis media was associated with prenatal exposure to DDE (RR=1.52), HCB (RR=1.49), and dieldrin (RR=1.26).
Hoyer *et al.* 1998, 2000, 2001 (*152-154*)	240 women with breast cancer and 477 matched controls that donated blood as part of the Copenhagen City Heart Study in 1976	Blood (serum) analyzed for mirex, dieldrin, aldrin, endrin, chlordane, HCH, heptachlor, HE, oxychlordane, trans-nonachlor, HCB, DDT, DDE, DDD, & 28 PCB congeners	Solid phase extraction with gas chromatography & electron capture detection	Questionnaire; lab analyses, relative risks and odds ratios calculated from conditional multiple logistic regression	A significant relationship between levels of dieldrin in blood and risk of breast cancer was found. A slight risk was also found between HCH, and risk of breast cancer was not associated with DDT, DDE, DDD or PCBs. The second study found that dieldrin also had significantly worsened breast cancer survival rates. The third study found that the results of the first study were most likely due to the fact that the women with high levels of dieldrin in their blood had estrogen receptor negative tumors which were larger and more spread out than the women with estrogen receptor positive tumors.

Table 2.1: Summary of the Applications of Organochlorine Biomarkers in Human Studies

Study	Study Subjects	Biomarker	Method of Detection	Study Methods	Study Findings
Fitzgerald *et al.* 1996, 1999 (*13,155*)	93 Native American men from a northern Great Lakes region	DDE & PCBs (68 congeners & total) in blood (serum)	Gas chromatography with electron capture detection	Consumption, wildlife & residential history; Lab tests; Breath tests; Regression analysis	Age, occupation, and fish consumption were found to predict serum PCB levels. DDE results not reported.
Helzlsouer *et al.* 1999 (*60*)	346 women that have been diagnosed with breast cancer & 346 matched controls without breast cancer who donated blood as part of the CLUE campaign(s)	DDT, DDE & PCBs in blood (serum & plasma)	Serum: solid phase extraction with gas chromatography & electron capture detection; Plasma: liquid/liquid extraction & adsorption chromatography	Questionnaire; Lab assays; Genotyping; Wilcoxon signed rank testing; Conditional Logistic regression	Total and congener-specific serum PCBs and total DDE were not found to cause an increased risk for breast cancer. Cases had a mean level of 1698.9 ppb of total lipid adjusted DDE and 735.3 ppb lipid adjusted PCBs, while controls had a mean of 1920.3 ppb DDE and 663.6 ppb PCBs. Adjustment for family history of breast cancer, BMI at age 20 or current, age at menarche, age at 1st birth, & duration of lactation did not alter odds ratios. Polymorphisms in the GSTM1, GSTT1, GSTP1, COMT, & CYP17 also did not alter the association between PCBs and breast cancer risk.
Kearney *et al.* 1999 (*156*)	232 fish license holders from a Great Lakes community	Organochlorines (14 PCB congeners, & 11 chlorinated pesticides) in blood (plasma)	Gas chromatography with electron capture detection	Great Lakes wildlife consumption & detailed questionnaire; Lab tests; Regression analysis	Great Lakes fish and waterfowl consumption were associated with increased levels of PCBs, DDE, & mirex. Breast feeding was associated with lower levels of PCBs, although it was not statistically significant.
Baris *et al.* 2000 (*157*).	22 non-Hodgkin's lymphoma patients	Blood serum analyzed for DDE and PCBs	Gas chromatography with electron capture detection	Lab analyses; Pretreatment & post treatment samples were compared using Pearson correlation coefficients & paired t-tests.	The study found that PCB and DDE levels in the blood of non-Hodgkin's lymphoma patients significantly decreased after chemotherapy treatment.

Table 2.1: Summary of the Applications of Organochlorine Biomarkers in Human Studies

Study	Study Subjects	Biomarker	Method of Detection	Study Methods	Study Findings
Fitzgerald *et al.* 1998, 2001 (*158,159*)	97 Native American women from a northern Great Lakes region and 154 Caucasian women from a control location. All had given birth 1 month prior	PCBs, DDE, mirex, and HCB concentrations in breast milk	Gas chromatography with electron capture detection	Questionnaire (consumption history, reproductive and pregnancy info); Lab tests; Multiple linear regression analysis	Mohawk mothers had significantly higher geometric mean levels of PCBs (0.602ppm), DDE, & mirex than control women (PCBs: 0.375 ppm) after controlling for previous breastfeeding, maternal age, alcohol consumption prior to pregnancy, & antibiotic use before pregnancy. These were found to be significant predictors of PCBs in breast milk. No difference was found in HCB levels. Organochlorine levels decreased over time in Mohawk women most likely due to local fish advisories.
Balluz *et al.* 2001 (*160*)	20 confirmed cases of systemic lupus erythematosus and 36 controls	Blood (serum) analyzed for β-HCH, dieldrin, HCB, HE, *trans*-nonachlor, mirex, oxychlordane, PCBs, DDT, DDE	Gas chromatography with electron capture detection	Lab analyses; odds ratios generated by conditional logistic regression	They did not find a statistically significant association between high levels of pesticides in the blood and disease status, but both cases and controls had elevated levels of organochlorines in their blood.
Belles-Isles *et al.* 2002 (*109*)	48 newborns from a subsistence fishing population; 60 newborns from a control population	Organochlorines (14 PCB congeners & 11 chlorinated pesticides & metabolites in cord blood (plasma)	High-resolution gas chromatography with dual Ni-63 electron-capture detectors	Maternal food frequency & lifestyle questionnaire; Cord blood analysis & immunological assays	PCB levels were significantly 3 times higher in subsistence group than controls. Subsistence group had a significant decrease in naïve helper T-cells, T-cell proliferation, & plasma IgM levels. A negative association was found between PCBs & DDE & T-cell clonal expansion. No association was found between natural killer cytolytic activities & PCBs.

Cantor *et al.* 2003 (*148*)	74 CLUE participants that have been diagnosed with non-Hodgkin lymphoma (NHL) & 147 matched controls without NHL	Chlordane, γ-HCH, β-HCH, transnonachlor, heptachlor, heptachlor epoxide, oxychlordane, dieldrin, HCB in blood (serum)	Serum: solid phase extraction with gas chromatography & electron capture detection	Questionnaire;
				Lab assays;
				Wilcoxon signed rank testing;
				Conditional Logistic regression
					No evidence of an association between these organochlorines and NHL risk were found.

2.2.7.2.1. Occupational Studies

There have been several occupational studies that have utilized biomarkers of organochlorine exposure. In one such study, several variables were used to predict serum PCB levels in workers at an electrical capacitor manufacturing plant. Acquavella *et al.* found duration of employment, cumulative occupational exposure, cumulative fish consumption, and cholesterol level to be significant risk factors for high levels of PCBs in the blood (*79*). In another study of electrical capacitor manufacturing workers, plasma PCB levels were measured to study the relationship between PCB exposure and dermatological health effects (*77*). Blood PCB levels in electrical workers have also been studied, and found to be associated with adverse effects in the liver (*76*). In addition, serum PCB levels have been used in conjunction with serum screening for oncogene proteins to study the possible increased risk for cancer in municipal workers who cleaned oil from old transformers in a study by Brandt-Rauf *et al.* (*150*). DDT's metabolite, DDE was measured in the adipose tissue of pesticide workers in Mexico, and the use of protective gear and recent weight loss were found to be effective in reducing exposure to DDT (*45*). Furthermore, serum HCH levels were found to be 12 times higher in exposed workers of an HCH manufacturing plant as compared to their controls in a study by Nigam *et al.*(*161*). A study by Sala *et al.* found that once workers retired from their jobs at an electrochemical factory, HCB levels in their blood declined (*3*).

2.2.7.2.2. Great Lakes Studies

Biomarker data on organochlorine exposure have also been used to study populations in the Great Lakes region. Because of their northern location, and the numerous industries along the shores of the Great Lakes and their tributaries, the region has become highly impacted by persistent organic pollutants, such as organochlorines. Studies have found that some large fish in the Great Lakes contain total DDT and PCB levels that exceed the tolerance level of five parts per million (ppm) set by the FDA for human consumption (*162*). There are many populations in that region that rely on subsistence fishing. Therefore, people living in the Great Lakes make ideal study populations. One study measured PCBs in the cord blood of mothers when they gave birth. Women who relied on subsistence fishing had PCB levels that were two to three times higher than the control group (*163*). In a similar study, eating fish and waterfowl from the Great Lakes region was associated with increased levels of PCBs and DDE in blood (*156*). In another study by Fitzgerald *et al.*, body burdens of PCBs and DDE were measured in the serum of men to determine if diet, residence, and occupation are associated with levels of PCBs and DDE in the blood (*13*).

2.2.7.2.3. CDC-NHANEs Studies

Information on organochlorine exposure in the general U.S. population has been collected by the Centers for Disease Control and Prevention (CDC), as a part of the Second National Health and Nutrition Examination Survey (NHANES II). Blood samples were taken from a representative population consisting of over 7,000 U.S. residents ranging from 12 to 74 years of age, and analyzed for organochlorines. Median DDT and o'p'-DDE concentrations were found to be below two ppb while the median p'p'-DDE blood level was 12.6 ppb (*4*). Using these data, researchers have reported that age, gender, and farming are associated with organochlorine levels in blood. In addition, race, income, and geographic location were found to be risk factors for some specific organochlorines (*80*). Another study measured organochlorines in the blood of a high risk population, dairy farm family members, and the results were compared to the NHANES II data since they are considered representative of the entire U.S. population (*105*).

2.2.7.2.4. CLUE Studies

In Washington County, Maryland, the Johns Hopkins Training Center for Public Health Research has been established to investigate chronic illnesses such as cancer and heart disease via large population-based epidemiological studies (*164*). The Center collected blood samples at two different time points from approximately one third of the Washington County population as part of the Campaign Against Cancer and Stroke (CLUE I) and Campaign Against Cancer and Heart Disease (CLUE II). Many of these samples were analyzed for organochlorines, and used in studies as biomarkers for organochlorine exposure. One study analyzed the association between organochlorine concentrations in blood and the risk of breast cancer. No statistically significant associations were found. However, the total DDE concentration levels in blood ranged from less than 6.93 to 80.33 nanograms per milliliter (ng/ml) unadjusted for lipids, and less than 1,017.19 to 10,795.91 ng/g lipid (*60*). Two studies by Rothman *et al.* and Cantor *et al.* used the blood samples analyzed for organochlorines to investigate whether organochlorine exposure was a risk factor for non-Hodgkin lymphoma. A strong dose-response relationship between blood PCB levels and the risk for non-Hodgkin lymphoma was found. They found total DDT and DDE mean levels to range from 8.7 to 49.1 ng/ml unadjusted for lipids, and 180 to 20,500 ng/g lipid (*131,148*). In addition, another study used the CLUE biomarker data to look at risks for prostate cancer. However, no associations were found, and the data were not published. This study also uses this blood organochlorine data, and therefore, more detailed information about the Center and its accomplishments are discussed in the next section.

2.3. CLUE I and II Study Design

2.3.1. Population

In 1962, a Johns Hopkins Training Center for Public Health Research was founded in Hagerstown, Washington County, Maryland (*164*). Among its many accomplishments, the Center sponsored the collection of two large blood specimen banks as part of the CLUE I and CLUE II campaigns. The campaigns were given their name from the slogan "give us a clue to cancer and heart disease." All adult residents of Washington County were asked to participate. Over 20,000 residents volunteered for CLUE I from August through November of 1974. These participants represented roughly one third of the Washington County population. In May through October of 1989, the CLUE II campaign recruited approximately one third of the population to participate as well. An estimated 25 percent of the CLUE I participants also participated in CLUE II (*131*). Participant ages ranged from 30 to 80 years old.

2.3.2. Data Collection

After signing a general consent form that allowed the Center to use their blood for research purposes, volunteers donated their blood in a number of trailers located throughout the County. At that time, brief questionnaires were administered to each of the blood donors. These questionnaires asked for a brief medical history, and information about medication use. Additional information about the participants has been obtained by the Center throughout the years. The Center sponsored two private Washington County censuses in 1963 and 1975, administered a food frequency questionnaire in 1989, and conducted a Women's Health Survey in 1995. In addition, CLUE II follow-up questionnaires were administered in 1996, 1998, and 2000. The information obtained from these efforts and evaluated in this study is summarized in Table 2.2 below.

Table 2.2: Summary of Informational Variables Obtained from Washington County Studies

Variable	CLUE I (1974)	1975 Census	CLUE II (1989)	Diet Study (1989)	Women's Health Survey (1995)
Age	X		X		
Gender	X		X		
Ethnicity	X		X		
Years of Schooling	X		X		
Marital Status	X		X		
Social Economic Status		X			
Height			X		X
Weight			X		X
Occupational exposure to pesticides					X
Smoking Status		X	X		
Water source and type		X			
Cholesterol level	X		X		
Diet: meat, fish, dairy, eggs, vegetables, fruit				X	
Breastfeeding history					X
Alcohol use				X	
Pesticide usage					X

The additional informational sources do not provide data on all of the blood donors from CLUE I and II. Instead, they provide links to information on only those donors that participated in that particular survey or questionnaire. In addition, there is a lack of information regarding potential exposure sources. For instance, details about occupation are limited for most of the CLUE participants. Finally, the early questionnaires did not obtain height and weight measurements. This information can be used to calculate body mass index, which may be a significant predictor of body burdens of lipophilic pollutants, such as organochlorines (*3,85,165-167*).

2.3.3. Laboratory Methods

As volunteers donated their blood in 1974 as part of the CLUE I campaign, the blood was drawn into 15 milliliter (ml) Vacutainer tubes. The blood was then refrigerated and allowed to clot before centrifugation. Two six ml aliquots of serum were removed and stored in –70 degrees Celsius (°C) temperatures. In 1989, during the CLUE II campaign, 20 ml of blood were collected into heparinized Vacutainer tubes. The tubes underwent centrifugation, and the plasma, buffy coat, and 1.8 ml of packed red blood cells were aliquotted and stored at -70 °C (*60,131,148*). These samples were stored at these temperatures until they were analyzed several years later. Research has shown that organochlorines are unlikely to degrade in frozen samples over time (*130,168*).

Two different laboratories were used to analyze the blood samples for organochlorines. The CDC laboratory in Atlanta, Georgia was used to analyze the blood samples for the breast cancer and non-Hodgkin lymphoma studies mentioned above.

Blood samples for the prostate cancer study were analyzed by Pacific Toxicology Laboratories in Los Angeles, California. Both labs were asked by the researchers to report true values even if they were below the official limit of detection for the method due to the theory that any reported values are more valid than interpolated values below the detection limit (*148*). Furthermore, Rothman *et al* reported that using one half of the detection limit instead of the real values had no effect on their analysis (*131*).

All the blood samples were analyzed for DDT, DDE, and 28 PCB congeners and reported as both unadjusted and adjusted for lipid content using the methods of Phillips *et al.* (*145,169*). Phillips *et al.* reported that PCB and DDE levels were over 20 percent higher in the blood of people that did not fast before the blood was taken (*169*). Therefore, adjusting for lipids controls for the fact that most participants did not fast before donating their blood. Less than half of the blood samples were also analyzed for alpha- and gamma-chlordane, oxychlordane, *trans*-nonachlor, heptachlor, heptachlor epoxide, HCB, mirex, dieldrin, aldrin, endrin, and beta- and gamma-HCH as part of breast and non-Hodgkin lymphoma cancer case control studies that took place in Washington County using the CLUE blood samples (*60,131*). Table 2.3 summarizes the organochlorine data from the three epidemiological studies that were available for this study as well as the published NHANES II data mentioned above.

Table 2.3: Summary of Available Blood Organochlorine Samples from Three CLUE Epidemiological Studies and a Summary of the NHANES II Blood Organochlorine Data

	Clue I				Clue II				NHANES Reference Background[*]			
Organochlorine	#	Range	Mean ± Standard Deviation	95% Confidence Interval	#	Range	Mean ± Standard Deviation	95% Confidence Interval	#	Range[†]	Mean	95% Confidence Interval
p,p DDE												
lipid adjusted (ng/g)	1289	0-46010	2643.7 ± 634	(2518.3, 2769.0)	708	0-10020	1347.2 ± 40.4	(1268.0, 1426.3)	1964	86-2020	297	(267, 330)
unadjusted (ng/ml)	1390	0-417	20.3 ± 0.53	(19.2, 20.3)	796	1-61.9	8.3 ± 0.25	(7.8, 8.8)				
p,p DDD												
lipid adjusted (ng/g)	470	0-164.8	0.68 ± 0.38	(-0.06, 1.42)	73	NA	NA	NA	NA	NA	NA	NA
unadjusted (ng/ml)	470	0-1.11	0.005 ± 0003	(-0.0003, 0.01)	73	NA	NA	NA				
o,p DDT												
lipid adjusted (ng/g)	470	9-1027	35.7 ± 3.6	(28.5, 42.8)	73	0-173.9	5.3 ± 2.8	(-0.12, 10.8)	1002	NA	<10.6	NA
unadjusted (ng/ml)	470	0-4.8	0.24 ± 0.02	(0.2, 0.57)	73	0-1.20	0.03 ± 0.02	(-0.004, 0.06)				
PCB congener 28												
lipid adjusted (ng/g)	714	0-302.2	62.7 ± 2.0	(58.7, 66.6)	73	0-110.8	9.7 ± 3.1	(3.6, 15.7)	1202	NA	<16.7	NA
unadjusted (ng/ml)	714	0-2.3	0.44 ± 0.01	(0.41, 0.47)	74	0-0.6	0.05 ± 0.02	(0.02, 0.09)				
PCB congener 52												
lipid adjusted (ng/g)	700	0-194.5	22.3 ± 1.2	(20.1, 24.6)	287	0-191.1	3.7 ± 1.2	(1.3, 6.1)	1248	NA	<6.4	NA
unadjusted (ng/ml)	700	0-1.0	0.15 ± 0.007	(0.14, 0.17)	287	0-1.2	0.02 ± 0.007	(0.008, 0.04)				
PCB congener 74												
lipid adjusted (ng/g)	756	0-919.2	68.7 ± 2.4	(64.0, 73.5)	102	0-310.6	56.7 ± 5.0	(47.0, 66.4)	1253	<6.4-30.0	<6.4	NA
unadjusted (ng/ml)	756	0-6.5	0.5 ± 0.02	(0.45, 0.51)	103	0-2.1	0.33 ± 0.03	(0.28, 0.39)				
PCB congener 101												
lipid adjusted (ng/g)	717	0-966	9.7 ± 1.7	(6.4, 12.9)	286	0-70.1	1.2 ± 0.42	(0.38, 2.0)	1260	NA	<13.2	NA
unadjusted (ng/ml)	717	0-6.8	0.07 ± 0.01	(0.05, 0.09)	286	0-0.44	0.007 ± 0.003	(0.002, 0.01)				
PCB congener 118												
lipid adjusted (ng/g)	1288	0-3631	233.7 ± 6.4	(211.1, 236.2)	650	0-930	98.1 ± 3.8	(90.6, 105.5)	1259	<6.4-43.6	<6.4	NA
unadjusted (ng/ml)	1291	0-841	2.6 ± 0.71	(1.2, 4.0)	689	0-7.0	0.60 ± 0.02	(0.56, 0.65)				

Table 2.3: Summary of Available Blood Organochlorine Samples from Three CLUE Epidemiological Studies and a Summary of the NHANES II Blood Organochlorine Data

	Clue I				Clue II				NHANES Reference Background*			
Organochlorine	#	Range	Mean ± Standard Deviation	95% Confidence Interval	#	Range	Mean ± Standard Deviation	95% Confidence Interval	#	Range†	Mean	95% Confidence Interval
PCB congener 138												
lipid adjusted (ng/g)	756	0-1883	150.2 ± 4.0	(142.4, 157.9)	322	0-555.4	85.9 ± 3.8	(78.4, 93.4)	1261	<21.1-72.8	<21.1	NA
unadjusted (ng/ml)	756	0-9.5	1.1 ± 0.03	(1.0, 1.1)	323	0-0.65	0.53 ± 0.02	(0.48, 0.58)				
PCB congener 156												
lipid adjusted (ng/g)	1066	0-857.8	33.4 ± 1.6	(30.4, 36.5)	390	0-574.1	17.4 ± 2.1	(13.3, 21.6)	1242	<6.4-17.5	<6.4	NA
unadjusted (ng/ml)	1070	0-828	1.3 ± 0.8	(-0.3, 2.9)	429	0-3.5	0.1 ± 0.01	(0.08, 0.1)				
PCB congener 170												
lipid adjusted (ng/g)	1271	0-278.9	39.0 ± 0.99	(37.1, 41.0)	643	0-166.3	31.2 ± 0.81	(29.6, 32.8)	1153	<8.9-33.9	<8.9	NA
unadjusted (ng/ml)	1274	0-10.1	0.30 ± 0.01	(0.28, 0.32)	682	0-1.7	0.19 ± 0.005	(0.18, 0.20)				
PCB congener 180												
lipid adjusted (ng/g)	1283	0-666.9	89.3 ± 2.3	(84.7, 93.9)	668	0-476.8	90.6 ± 2.5	(85.8, 95.5)	1257	<14.5-83.8	<14.5	NA
unadjusted (ng/ml)	1287	0-624	1.2 ± 0.49	(0.21, 2.1)	707	0-3.8	0.56 ± 0.01	(0.53, 0.59)				
PCB congener 183												
lipid adjusted (ng/g)	752	0-104.9	11.3 ± 0.57	(10.2, 12.4)	290	0-62.2	2.5 ± 0.47	(1.5, 3.4)	1260	NA	<6.4	NA
unadjusted (ng/ml)	752	0-0.92	0.08 ± 0.004	(0.07, 0.09)	291	0-0.42	0.02 ± 0.003	(0.01, 0.02)				
PCB congener 187												
lipid adjusted (ng/g)	757	0-344.4	14.3 ± 0.92	(12.5, 16.1)	318	0-166	10.9 ± 1.2	(8.4, 13.3)	1263	<6.4-25.9	<6.4	NA
unadjusted (ng/ml)	757	0-2.6	0.11 ± 0.007	(0.09, 0.12)	319	0-1.1	0.07 ± 0.008	(0.05, 0.08)				
PCB congener 194												
lipid adjusted (ng/g)	757	0-207.7	9.6 ± 0.64	(8.4, 10.9)	103	0-84.5	19.1 ± 1.9	(15.5, 22.7)	NA	NA	NA	NA
unadjusted (ng/ml)	757	0-1.8	0.07 ± 0.005	(0.06, 0.08)	104	0-0.57	0.11 ± 0.01	(0.09, 0.14)				
PCB congener 195												
lipid adjusted (ng/g)	705	0-2370	7.5 ± 3.4	(0.90, 14.1)	284	0-61.8	7.4 ± 0.53	(6.4, 8.5)	NA	NA	NA	NA
unadjusted (ng/ml)	705	0-20.2	0.06 ± 0.03	(0.002, 0.11)	284	0-0.45	0.05 ± 0.004	(0.04, 0.06)				

Table 2.3: Summary of Available Blood Organochlorine Samples from Three CLUE Epidemiological Studies and a Summary of the NHANES II Blood Organochlorine Data

	Clue I				Clue II				NHANES Reference Background*			
Organochlorine	#	Range	Mean ± Standard Deviation	95% Confidence Interval	#	Range	Mean ± Standard Deviation	95% Confidence Interval	#	Range†	Mean	95% Confidence Interval
PCB congener 201												
lipid adjusted (ng/g)	752	0-155.4	11.2 ± 0.76	(9.7, 12.7)	319	0-157.1	14.0 ± 1.2	(11.7, 16.3)	NA	NA	NA	NA
unadjusted (ng/ml)	752	0-1.6	0.08 ± 0.006	(0.07, 0.1)	320	0-1.1	0.09 ± 0.008	(0.07, 0.10)				
PCB congener 203												
lipid adjusted (ng/g)	757	0-314.1	11.9 ± 1.1	(9.8, 14.0)	321	0-240	16.5 ± 1.5	(13.6, 19.4)	NA	NA	NA	NA
unadjusted (ng/ml)	757	0-2.7	0.09 ± 0.008	(0.07, 0.10)	321	0-1.6	0.10 ± 0.01	(0.09, 0.12)				
PCB congener 206												
lipid adjusted (ng/g)	756	0-2159	10.3 ± 2.9	(4.6, 16.0)	112	0-148	29.9 ± 2.7	(24.6, 35.2)	NA	NA	NA	NA
unadjusted (ng/ml)	756	0-18.4	0.08 ± 0.03	(0.03, 0.13)	113	0-0.77	0.18 ± 0.02	(0.15, 0.21)				
Total of PCB congeners: 105, 146, 153												
lipid adjusted (ng/g)	1285	0-3125	273.8 ± 6.6	(261.0, 286.7)	669	0-1251	202.5 ± 6.0	(190.7, 214.3)	NA	NA	NA	NA
unadjusted (ng/ml)	1290	0-62.6	2.1 ± 0.08	(2.0, 2.3)	708	0-13.0	1.3 ± 0.04	(1.2, 1.3)				
Total of PCB congeners: 105, 118, 146, 153, 156, 170, 180												
lipid adjusted (ng/g)	1063	34.2-8022	643.5 ± 18.1	(608.1, 679.0)	384	0-2794.0	418.0 ± 19.8	(379.2, 456.9)	NA	NA	NA	NA
unadjusted (ng/ml)	1065	0.2-934.3	6.0 ± 0.94	(4.2, 7.9)	423	0-29.1	2.7 ± 0.13	(2.4, 2.9)				
Total of PCB congeners: 105, 118, 146, 153, 170, 180												
lipid adjusted (ng/g)	NA	NA	NA	NA	666	0-25.5	2.6 ± 0.08	(2.5, 2.8)	NA	NA	NA	NA
unadjusted (ng/ml)	NA	NA	NA	NA	627	0-2454	423.7 ± 12.1	(400.0, 447.5)				

Table 2.3: Summary of Available Blood Organochlorine Samples from Three CLUE Epidemiological Studies and a Summary of the NHANES II Blood Organochlorine Data

	Clue I				Clue II				NHANES Reference Background*			
Organochlorine	#	Range	Mean ± Standard Deviation	95% Confidence Interval	#	Range	Mean ± Standard Deviation	95% Confidence Interval	#	Range†	Mean	95% Confidence Interval
Total of PCB congeners: 28, 52, 74, 101, 105, 118, 138, 146, 153, 156, 170, 180, 183, 187, 194, 195, 201, 203, 206												
lipid adjusted (ng/g)	697	108.1-6162	747.4 ± 19.1	(709.9, 784.8)	72	17.6-3126	557.0 ± 58.8	(441.7, 672.2)	NA	NA	NA	NA
unadjusted (ng/ml)	697	0.68-41	5.3 ± 0.13	(5.0, 5.5)	73	0.09-20.7	3.1 ± 0.35	(2.5, 3.8)				
alpha chlordane												
lipid adjusted (ng/g)	694	0-158	5.1 ± 0.50	(4.1, 6.1)	73	0-21.7	0.91 ± 0.43	(0.06, 1.8)	NA	NA	NA	NA
unadjusted (ng/ml)	696	0-1.0	0.04 ± 0.003	(0.03, 0.04)	73	0-0.009	0.004 ± 0.002	(0.0004, 0.008)				
gamma chlordane												
lipid adjusted (ng/g)	466	0-616.8	29.4 ± 3.1	(23.3, 35.5)	73	0-80.0	2.1 ± 1.3	(-0.3, 4.6)	NA	NA	NA	NA
unadjusted (ng/ml)	467	0-4.6	0.20 ± 0.02	(0.16, 0.24)	73	0-0.55	0.01 ± 0.009	(-0.003, 0.03)				
Oxychlordane												
lipid adjusted (ng/g)	694	0-342.4	54.5 ± 2.2	(50.1, 58.9)	72	0-493.9	57.2 ± 9.6	(38.5, 76.0)	998	<7.4-47.7	<7.4	NA
unadjusted (ng/ml)	696	0-2.6	0.39 ± 0.02	(0.36, 0.43)	73	0-2.5	0.33 ± 0.05	(0.22, 0.43)				
trans Nonachlor												
lipid adjusted (ng/g)	694	0-360.2	36.7 ± 1.5	(33.7, 39.6)	282	0-414.1	53.1 ± 2.7	(47.7, 58.5)	1269	<7.5-82.2	20.8	(19.2, 22.6)
unadjusted (ng/ml)	697	0-2.7	0.27 ± 0.01	(0.25, 0.30)	287	0-2.5	0.33 ± 0.02	(0.30, 0.37)				
Heptachlor												
lipid adjusted (ng/g)	693	0-162.2	20.3 ± 1.1	(18.2, 22.4)	73	0-7.6	0.2 ± 0.12	(-0.04, 0.44)	NA	NA	NA	NA
unadjusted (ng/ml)	696	0-0.96	0.14 ± 0.007	(0.12, 0.15)	73	0-0.04	0.001 ± 0.0006	(-0.0002, 0.002)				

Table 2.3: Summary of Available Blood Organochlorine Samples from Three CLUE Epidemiological Studies and a Summary of the NHANES II Blood Organochlorine Data

	Clue I				Clue II				NHANES Reference Background*			
Organochlorine	#	Range	Mean ± Standard Deviation	95% Confidence Interval	#	Range	Mean ± Standard Deviation	95% Confidence Interval	#	Range[†]	Mean	95% Confidence Interval
Heptachlor Epoxide												
lipid adjusted (ng/g)	695	0-632.3	58.1 ± 2.9	(52.5, 63.7)	73	0-708.7	31.4 ± 11.4	(9.0, 53.8)	951	<7.5-27.1	<7.5	NA
unadjusted (ng/ml)	697	0-4.4	0.42 ± 0.02	(0.38, 0.46)	73	0-3.6	0.18 ± 0.06	(0.06, 0.31)				
Hexachlorobenzene												
lipid adjusted (ng/g)	227	1.6-171.9	25.2 ± 1.1	(23.0, 27.4)	282	0-107.2	36.9 ± 0.96	(35.0, 38.8)	1111	NA	<60.5	NA
unadjusted (ng/ml)	227	0.01-1.7	0.19 ± 0.009	(0.17, 0.20)	283	0-0.58	0.23 ± 0.006	(0.22, 0.24)				
Mirex												
lipid adjusted (ng/g)	692	0-884.7	4.7 ± 1.5	(1.8, 7.6)	283	0-261.4	3.3 ± 1.3	(0.67, 5.9)	1194	NA	<7.5	NA
unadjusted (ng/ml)	694	0-4.9	0.03 ± 0.009	(0.02, 0.05)	283	0-1.8	0.02 ± 0.009	(0.004, 0.04)				
Aldrin												
lipid adjusted (ng/g)	695	0-207.1	30.4 ± 1.3	(27.8, 32.9)	73	0-39.4	2.2 ± 0.9	(0.5, 3.8)	NA	NA	NA	NA
unadjusted (ng/ml)	696	0-1.1	0.20 ± 0.008	(0.19, 0.22)	73	0-0.2	0.01 ± 0.005	(0.003, 0.02)				
Dieldrin												
lipid adjusted (ng/g)	693	0-756.3	101.2 ± 2.8	(95.7, 106.6)	NA	NA	NA	NA	NA	NA	NA	NA
unadjusted (ng/ml)	696	0-5.6	0.71 ± 0.02	(0.67, 0.75)	NA	NA	NA	NA				
Endrin												
lipid adjusted (ng/g)	692	0-603.8	49.7 ± 3.7	(42.4, 57.0)	NA	NA	NA	NA	NA	NA	NA	NA
unadjusted (ng/ml)	694	0-3.0	0.33 ± 0.02	(0.28, 0.37)	NA	NA	NA	NA				
beta Hexachlorocyclohexane												
lipid adjusted (ng/g)	693	0-2042	107.9 ± 4.3	(99.5, 116.2)	282	0-378.2	54.7 ± 2.9	(49.1, 60.3)	1240	<4.8-73.4	10.9	(10.1, 11.7)
unadjusted (ng/ml)	696	0-14.7	0.79 ± 0.03	(0.73, 0.85)	283	0-1.9	0.34 ± 0.02	(0.31, 0.38)				

Table 2.3: Summary of Available Blood Organochlorine Samples from Three CLUE Epidemiological Studies and a Summary of the NHANES II Blood Organochlorine Data

	Clue I				Clue II				NHANES Reference Background*			
Organochlorine	#	Range	Mean ± Standard Deviation	95% Confidence Interval	#	Range	Mean ± Standard Deviation	95% Confidence Interval	#	Range†	Mean	95% Confidence Interval
gamma Hexa-chlorocyclohexane												
lipid adjusted (ng/g)	696	0-121.4	0.95 ± 0.34	(0.29, 1.6)	283	0-184.8	8.0 ± 0.75	(6.5, 9.4)	1139	NA	<7.5	NA
unadjusted (ng/ml)	697	0-0.71	0.007± 0.002	(0.002, 0.01)	283	0-0.94	0.05 ± 0.004	(0.04, 0.06)				

*NHANES reference group consists of age group 20 and over 20 yrs old in 1999-2000

† NHANES range reported as 10th percentile of data - 95th percentile

NA: Not Applicable/ Not Measured in Study

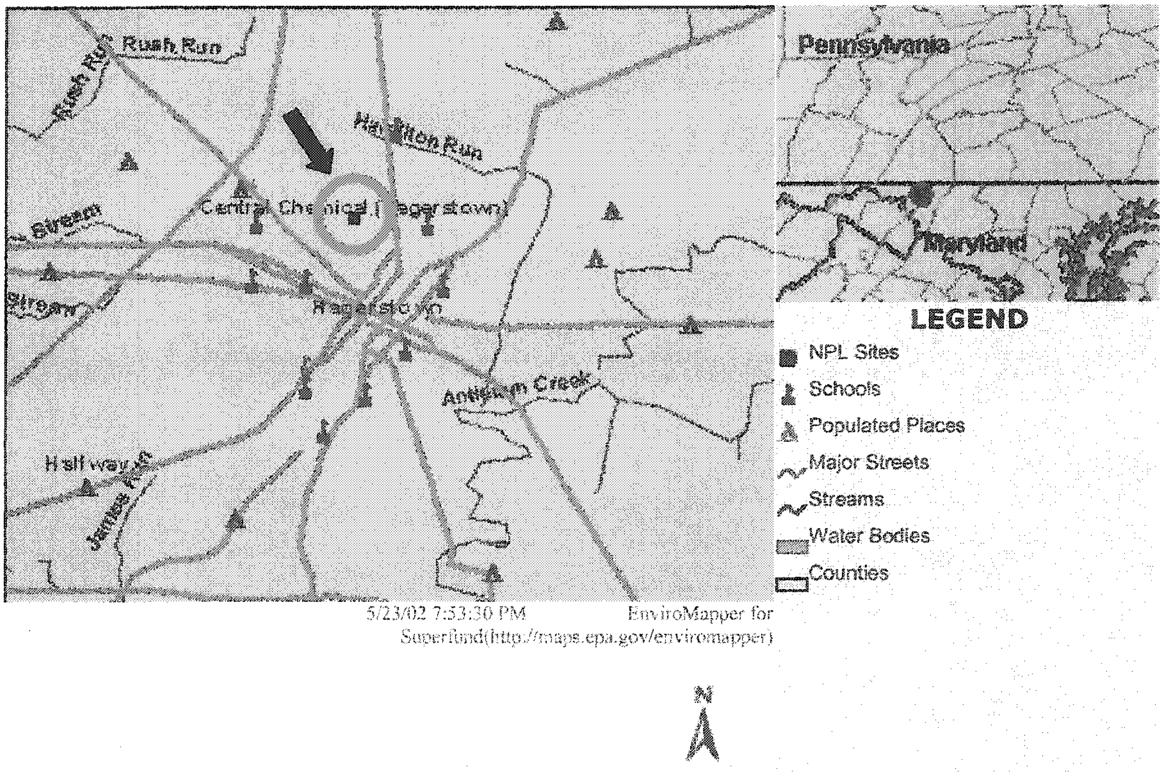
A detailed quality assurance/quality control program was followed for all analyses. Samples were pooled and analyzed to test for interset variability, and duplicate samples were used to test for intraset variability. Coefficients of between set variation ranged from 6.7 to 20.0 percent, and within set coefficients of variation ranged from 8.4 to 18.0 percent (*60,131,148*).

One source of variability in the blood organochlorine results may be due to the fact that the organochlorine concentrations measured in serum in 1974 while 1989 concentrations were measured in plasma. Plasma is whole blood, while serum is plasma minus the clotting agents, which may interfere with analytical techniques. Although this issue needs further exploration, its estimated that the difference between the two concentration measures, when adjusted for lipids, is likely to be small when compared the difference that should be seen between the two time points (*60*). This issue should not generate great variability in this research since each time period is analyzed separately, and only overall results are compared. Therefore, for the purposes of this study, the serum/plasma samples are referred to collectively as blood samples.

2.4. Superfund Site

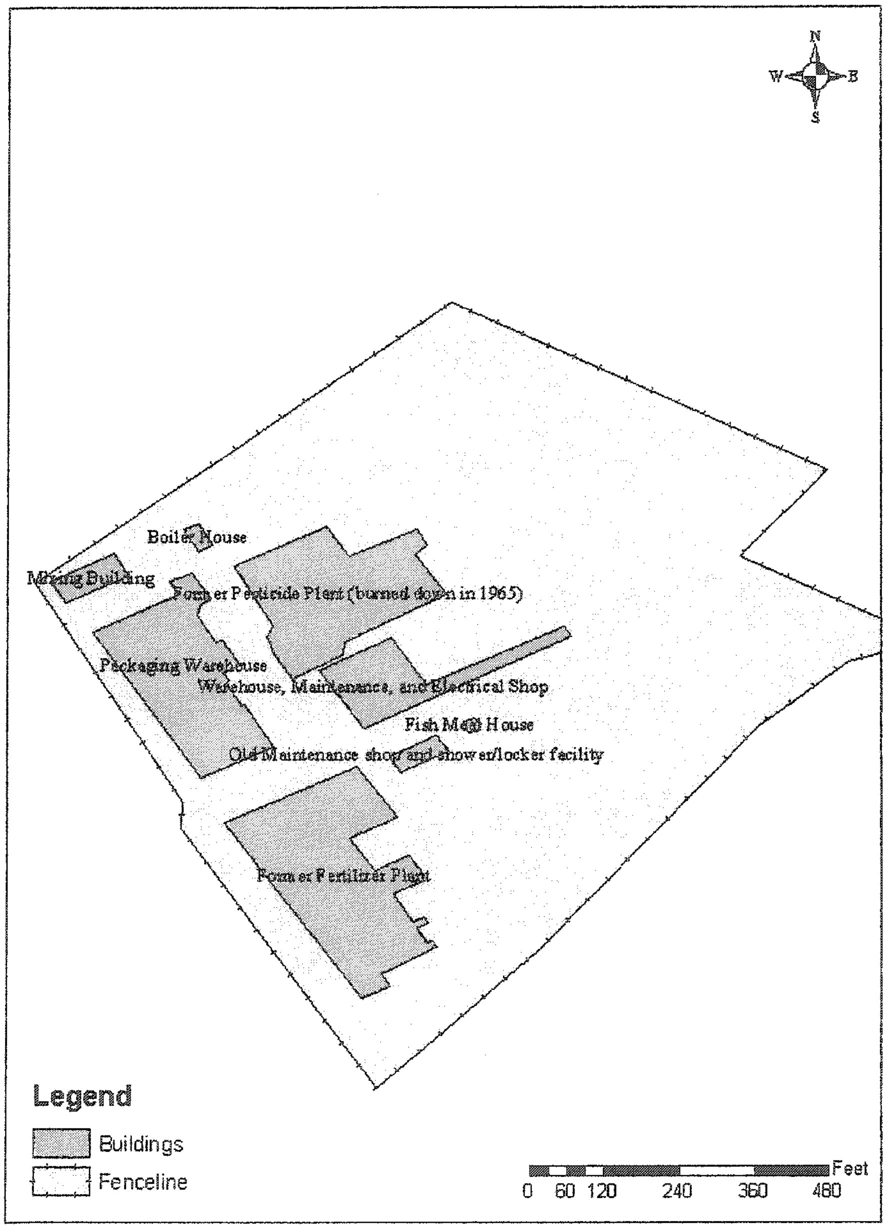
In the early 1930's, the Central Chemical Corporation (CCC) built a 19-acre facility in Hagerstown, Maryland. Figures 2.4 and 2.5 depict a map of the area and the layout of the site, respectively. The site blended pesticides, such as DDT, Sevin, chlordane, Guthion, Daconil, and Omite with inert materials, such as clay. In 1965, a fire burned the main pesticide blending building on the premises, and most of the pesticide operations at the site ceased. However, smaller scale pesticide blending continued to take place in another building on the site for three more years. In 1968, the facility terminated all of its pesticide operations. The facility converted to a fertilizer plant, blending ammonia solutions with other materials until 1984. In 1984, the site became inactive (*170*). The site contained only one electrical substation, owned by the City of Hagerstown, that used PCBs (*8*).

Figure 2.4: Map of the Superfund site location in the City of Hagerstown and surrounding area



5/23/02 7:53:30 PM EnviroMapper for Superfund(http://maps.epa.gov/enviromapper)

Figure 2.5: Map of the CCC Superfund site



In 1970, Maryland Department of Water Resources (MDWR) personnel found a small dump area outside of the plant premises that consisted of stagnant water, bags marked as insecticides, and a "suspicious" dust. Under the direction of the Washington County Health Department, the dump was covered with two feet of soil. In 1987, a second buried dumpsite was found on adjacent property when a trench was excavated for a sewer line. The site contained pesticides, naphthalene, and volatile organic compounds (VOCs) (*170*).

Washington County residents began complaining about the CCC site as early as the 1960's when they claimed pesticide odors were migrating off site. The Maryland State Health Department found levels of the pesticide, Guthion, in the air, but not at levels they believed to be hazardous. In 1970, there were more complaints from nearby residents that an air stack on the site was emitting dust and smoke. In 1972, the plant opted to stop using the fertilizer granulator that was responsible for the dust and smoke, instead of installing emission control equipment to comply with State Air Regulations (*170*).

In addition, specific organochlorines from the CCC site, such as benzene hexachloride, aldrin, chlordane, dieldrin, methoxychlor, DDT, DDE, and DDD, were found to have migrated off site in water. In 1976, water samples from the nearby Antietam Creek indicated that some pesticides, such as DDT, and lead were being transported in the creek from the Hagerstown area. In order to find the source of this contamination, samples were taken from the CCC site, and high levels of DDT, lead, and arsenic were found. More samples taken in 1993 by the Maryland Department of the Environment (MDE) confirm the presence and migration of pesticides and organic constituents in both the soil and groundwater at the site. Furthermore, levels of DDT and its derivatives, DDE and DDD, were found in the edible filet portions of fish caught in Antietam Creek (*170*).

A list of organochlorines that were evaluated in this thesis and whether or not they are associated with the CCC Superfund site according to EPA reports is found in Table 2.4 below.

Table 2.4: Organochlorines Evaluated and Their Relationship to the Superfund Site

Organochlorine	Used at Site	
	Yes	No
Aldrin	X	
Dieldrin*	X	
Endrin		X
α-Chlordane	X	
γ-Chlordane	X	
Oxychlordane*	X	
trans-Nonachlor		X
Heptachlor		X
Heptachlor epoxide*		X
Hexachlorobenzene	X	
β-Hexachlorocyclohexane		X
γ-Hexachlorocyclohexane (lindane)		X
Mirex		X
o,p'DDT; p,p'DDT;	X	
p,p'DDE*	X	
p,p'DDD*	X	
PCBs		X

* Serum/plasma metabolite of parent compound

This contamination prompted the EPA to propose the site to be on the National Priorities List (NPL) for clean up in 1996. Next, the EPA screened and scored the site based on its potential to threaten human health or the environment using the Hazard Ranking System (HRS). On September 25, 1997, the CCC site became a Superfund site.

The site is covered by the Comprehensive Environmental Response Compensation and Liability Act (CERCLA). Currently, the site is in the Remedial Investigation/Feasibility Study (RI/FS) stage in which potential responsible parties have developed a work plan to characterize the extent of contamination on the site (*171*). Refer to Table 2.5 for a detailed timeline of events regarding the CCC Superfund site. Some of the fourteen potentially responsible parties include: Allied Signal, Central Chemical, FMC, Novartis, Olin, Shell Oil, Union Carbide, and Wilmington Securities. After the RI/FS is completed, a Record of Decision will be made as to what clean up alternatives will be used on the site, and the site will enter the Remedial Design/Remedial Action phase of the Superfund process during which the clean up process will begin. Once the clean up is completed, the EPA will continuously monitor the site to make sure it is no longer threatening human health or the environment. Eventually, it will be taken off the NPL (*172*).

Table 2.5: Timeline of Events for the Central Chemical Company Superfund Site	
Year	**Event**
1930's	19-acre facility built in Hagerstown to blend pesticides with inert materials (clay)
1960's	Residents complained of pesticide odors
1965	Fire burned down main pesticide operation building
1968	Facility ceased pesticide production, and converted to fertilizer plant, blending and drying ammonia solutions
1970	MDWR found small dump offsite consisting of stagnant water and bags of insecticides
	Residents complained air stack was emitting dust and smoke
1973	Pesticides were found in downstream water of Antietam Creek and on site
1976	MDE collected soil and sediment samples
1977	CCC hired Baker and Wibberly to collect soil, groundwater, and surface water samples
1984	The CCC site became inactive
1987	Another dump was found on adjacent property containing pesticides and VOCs
	EPA put the CCC site on CERCLIS
1988	CCC hired Roy F. Weston, Inc. to collect soil and groundwater samples, as well as to conduct an electromagnetic and ground penetrating radar survey
1989	MDE completed a Screening Site Investigation, and collected soil, sediment, and groundwater samples
1991	MDE completed the Preliminary Assessment of the CCC site
1992	EPA conducted a Removal Assessment of the CCC site. Soil, groundwater, and wipe samples were collected.
1993	MDE completed an Expanded Site Investigation. Soil, surface water, sediment, groundwater and wipe samples were collected.
1994	Agency for Toxic Substances and Disease Registry (ATSDR) recommended a sampling strategy, and EPA collected soil and fish samples
	Presence of off-site migration of pesticides in soil, groundwater, and edible fish was confirmed
1996	EPA proposed the CCC site for the NPL
1997	EPA scored the level of contamination on site by completing the HRS
	CCC was put on the NPL
2002	CCC site was in the RI/FS stage where the 7 potential responsible parties submitted proposed work plans for site cleanup.
2003	Potential responsible parties completed the RI/FS work plan outlining what is necessary to determine the extent of contamination.
	Phase I of the work plan was completed. Groundwater monitoring wells were installed, and environmental samples were taken.
	City of Hagerstown put together a community-based report stating that the city preferred the land be redeveloped as light industry or commercial office park with a natural buffer area once the site is cleaned up.
2004	Phase I work plan results will be interpreted and remaining field work (phase II) is to be completed.
2005	RI/FS is expected to be completed, and EPA will issue a proposed cleanup plan

OVERVIEW OF RESEARCH METHODS

3. Research Methods

The goal of this research was to determine the geographic distribution characteristics of human biomarkers of exposure to persistent organochlorine chemicals in a community with a potential source. In order to accomplish this goal, it was first necessary to determine whether or not a former pesticide facility located within the population was a potential source of exposure to surrounding residents. The next step was to gain a better understanding of the potential risk factors for exposure to organochlorines including both established individual potential risk factors and additional spatially dependent potential risk factors. These steps were accomplished with regression analyses. First, soil organochlorine concentrations were predicted in and around the site using geostatistical regression techniques. In the second step, generalized least squares regression was used to model the concentration level of each specific organochlorine compound in the blood of Washington County, Maryland residents collected in the CLUE I and II studies in 1974 and 1989, respectively. Participant and residence information came from the multiple information sources collected by the Johns Hopkins Training Center for Public Health Research, and was used as covariates in these models. Once both the 1974 and 1989 regression models were developed, the influence of spatial variables and spatial dependence in the data were evaluated. Finally, the robustness of these models to positional inaccuracies in residential location was tested. These processes are outlined in more detail in this section. However, an introduction to geographic information systems (GIS) and a description of the data collection processes are provided first.

3.1. Geographic Information Systems

GIS are computer-based systems that serve as powerful tools to input, store, retrieve, manipulate, descriptively analyze, and output data that have geographic characteristics (*2,20*). GIS provides methods for the user to query data, study the relationship between any selected geographical or spatial variables, and generate a comprehensive spatial database. In addition, GIS can serve as a relational database by providing a common format to link specific data to additional geographical or environmental data collected for other purposes in the same geographic area. Each of these databases can be mapped, and layered atop of one another in a process called overlaying. The U.S. Geological Survey (USGS) has established procedures to properly access, link, and document these data bases (*2*).

GIS has been used extensively in the environmental, healthcare, and public health fields. GIS can be used to map environmental contamination data, track air emission, spill plumes, manage health care organizations, and monitor diseases (*12-19,173-187*). For example, Glass *et al.* used GIS to study environmental risk factors for lyme disease (*18*). More pertinent to the proposed study, Fitzgerald *et al.* used GIS to organize and map over 8,000 measurements of environmental organochlorine contamination data from 40 different documents to study the extent of environmental contamination in the St. Lawrence River region (*13*). In this study, GIS is used to combine environmental data, as well as information from CLUE studies and questionnaires, such as residential location so the spatial distribution of organochlorines in human blood can be characterized as described in more detail below.

3.2. Data Collection

3.2.1. Environmental Data

Almost 4,000 environmental samples have been collected at the Central Chemical Company (CCC) Superfund site and analyzed for 1,1,1-trichloro-2,2-bis(*p*-chlorophenyl)ethane (DDT), 1,1-dichloro-2,2-*bis*(chlorophenyl) ethylene (DDE), and 1,1-dichloro-2,2-*bis*(p-chlorophenyl) ethane (DDD) and other compounds at the request of the Maryland Department of the Environment (MDE). These data include soil, surface water, groundwater, drinking water, air, river sediment, and wildlife samples. The results of these investigations were published in reports which were attained via the Freedom of Information Act (*8-10,170,188*). Within these reports, contamination data were presented along with survey-type maps of the site that included drawings of buildings and other landmarks. Locations of sampled data were marked on these maps and could be cross-referenced to the concentration value using the corresponding unique sample identifier. For this thesis, all environmental samples were combined into one large electronic database. The GIS process of on-screen digitization was used to determine the approximate coordinates of 110 locations from which soil samples had been collected and analyzed for DDE in the 1990's since these samples represented the most complete set taken within a five-year period. Geocoding was then used to link the soil concentration levels from the database to the collection locations. These data were then mapped and used to predict levels of DDE in soil at unsampled areas including areas undergoing residential development. These processes are described in more detail in the first manuscript entitled, "GIS and Geostatistics: Tools for Characterizing Environmental Contamination."

3.2.2. Address Data

In order to determine the residential location of the study subjects, address information was obtained on all individuals. Participant street addresses and zip codes of the participants were collected as part of the CLUE campaigns and accessed through the Johns Hopkins Training Center for Public Health Research. Since CLUE I addresses were not in electronic form, many were obtained electronically through the 1975 private census data. In order to ensure that these addresses were the same, each electronic address was double-checked against the actual CLUE survey for accuracy. If the subject had moved between the 1974 CLUE campaign, and the 1975 private census, the 1974 address was entered in place of the 1975 address obtained via the census. For those that did not participate in the census, the handwritten addresses on the original CLUE survey were entered into the database. Accuracy was verified when a ten percent random sample of the entered addresses was checked. CLUE II addresses were already in electronic form. Apartment numbers were removed, obvious spelling mistakes were fixed, and the address data were re-organized so that they may be brought into GIS to determine their locations.

However, in the early 1990's, Washington County changed its address system, and renamed several roads and renumbered the street addresses as part of an emergency response improvement. Specifically, road names were changed so that no two streets shared the same name, rural routes were replaced with a street name, and house numbers outside city limits were given new numbers of at least five digits. Therefore, County directories from both before the change and after had to be used to deduct the current valid address that corresponded with the CLUE address. Furthermore, a new zip code was added to Washington County to account for the growing population.

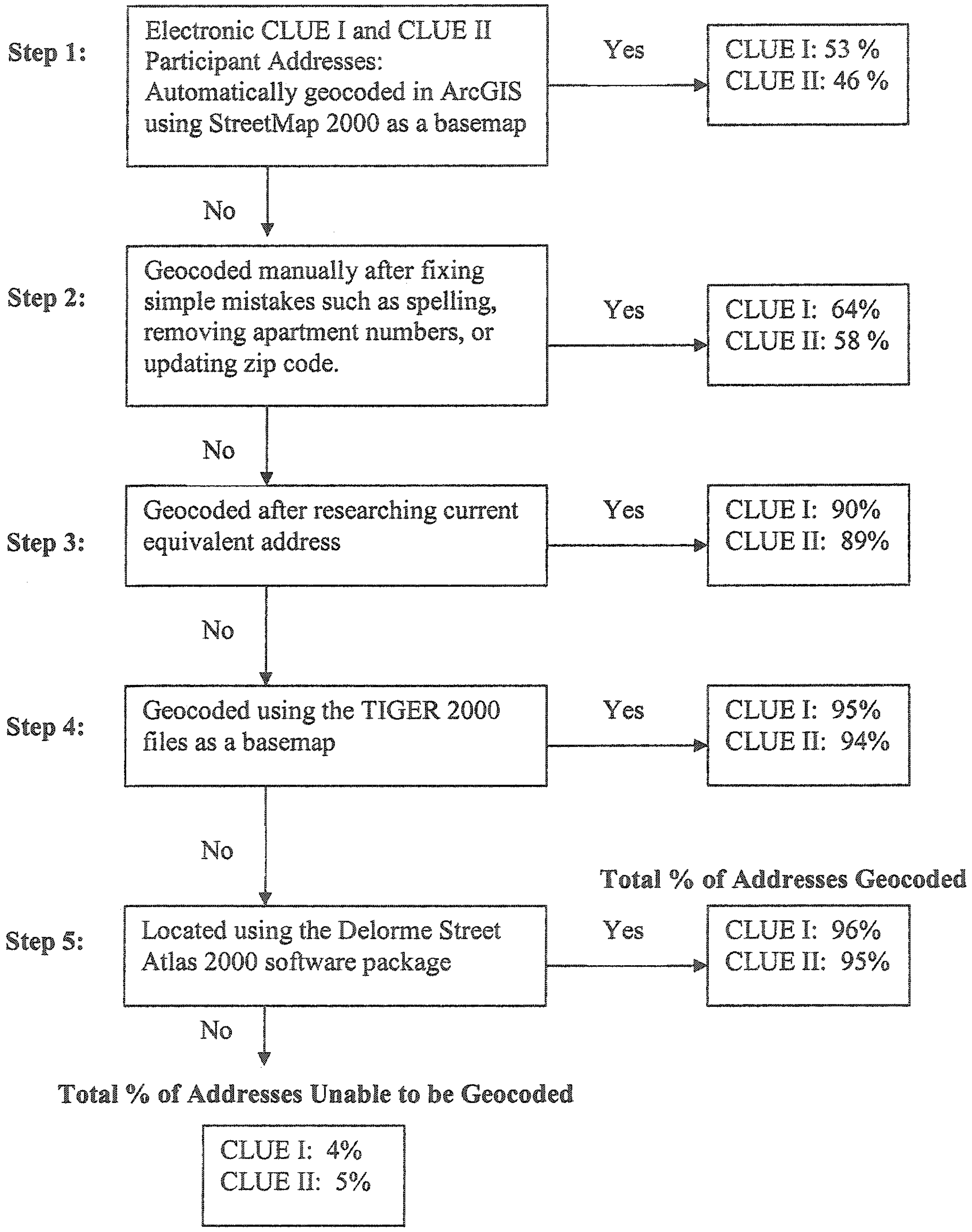
Results of the address analysis indicated that approximately 50 percent of the CLUE I and CLUE II addresses were current valid addresses as they were obtained. An additional approximate 10 percent of the addresses from each CLUE campaign became valid after minor fixes, such as spelling errors, or accounting for the zip code change. The remaining addresses had undergone major changes, and the new corresponding address had to be researched using the County directories as mentioned above.

Once as many of the addresses were converted to a valid current form as possible, they were brought into ArcGIS (Environmental Systems Research Institute, Inc.) to find their geographical location through geocoding. Geocoding is the GIS process by which an attribute or characteristic can be linked to a geographical location. Most often, this is accomplished using ordinary street addresses that GIS matches to digital basemaps within its system, providing, for example, a set of longitude and latitude coordinates. With this information, GIS can serve as a relational database by providing a common format to link specific data at the location to additional geographical or environmental data collected in the same geographic area (*2*). Such spatial databases provide the necessary input to the various methods of spatial statistics available.

Two basemaps were used in this study to link addresses to a geographical location (*177*). They are StreetMap 2000 (Environmental Systems Research Institute, Inc.) and the United States Census Bureau 2000 Topologically Integrated Geographic Encoding and Referencing (TIGER) street maps. Approximately 91 percent of CLUE I and 88 percent of CLUE II addresses were geocoded using the StreetMap 2000 basemap. The

TIGER file basemap located an additional five percent of the addresses. The Delorme Street Atlas software package (Delorme) was then used in an attempt to locate the remaining addresses. This approach was recommended by McElroy *et al.* to achieve maximum geocoding success (*177*). In total, approximately 96 percent of the CLUE I and 95 percent of the CLUE II addresses were able to be geographically located. Figure 3.1 is a visual display of this geocoding process. The geocoded latitude (y coordinate) and longitude (x coordinate) of each address location was stored in a database along with the subject ID code. In situations where several CLUE participants live in one location, such as an apartment building, one foot in both the x and y directions was added in increments to the residential location to make sure all participants had a unique location, and no two participants were grouped into one when mapping.

Figure 3.1: Flow chart of the geocoding process used to locate the residential location of CLUE participants whose blood was analyzed for organochlorines


Step 1:
Electronic CLUE I and CLUE II Participant Addresses: Automatically geocoded in ArcGIS using StreetMap 2000 as a basemap
Yes
CLUE I: 53 %
CLUE II: 46 %
No
Step 2:
Geocoded manually after fixing simple mistakes such as spelling, removing apartment numbers, or updating zip code.
Yes
CLUE I: 64%
CLUE II: 58 %
No
Step 3:
Geocoded after researching current equivalent address
Yes
CLUE I: 90%
CLUE II: 89%
No
Step 4:
Geocoded using the TIGER 2000 files as a basemap
Yes
CLUE I: 95%
CLUE II: 94%
No
Total % of Addresses Geocoded
Step 5:
Located using the Delorme Street Atlas 2000 software package
Yes
CLUE I: 96%
CLUE II: 95%
No
Total % of Addresses Unable to be Geocoded
CLUE I: 4%
CLUE II: 5%

3.2.3. Geocoding Positional Bias Data

Geocoded locations from street addresses are not 100 percent accurate. GIS basemaps contain roads, which are represented as line segments, and road intersections. Geocoded coordinates are obtained by approximating numbered addresses in proportion to the street segments between intersections. For example, the geocoded coordinate for 125 Main Street would be the location representing one quarter of the way between the 100 and 200 blocks of Main Street, assuming the basemap contains these block intersections. The distance along the road between the true location and the geocoded coordinate is one source of geocoding positional bias. The distance residences are from the street is another source of this bias.

In order to quantify biases associated with geocoding addresses in GIS, a random sample of 163 study participant addresses, from a total of 1,323 were chosen from rural and urban Washington County zip codes, and those addresses were visited. The distance between the geocoded coordinates of the address and the coordinates taken with a global positioning system (GPS) in front of the residence was measured. Furthermore, a laser distance meter was used to measure the distance between the street and the residence. More details regarding this geocoding positional bias data collection can be found in the third manuscript entitled, "Geocoding Positional Bias and its Effect in Analyzing the Spatial Distribution of Blood Dieldrin Levels"

3.2.4. Blood Data

This research uses samples from both the CLUE I and CLUE II blood banks that have already been analyzed for organochlorines for use in the breast cancer, prostate cancer, and non-Hodgkin lymphoma studies that were mentioned in the background section. These data were accessible after receiving Committee on Human Research (CHR) approval (CHR # H.18.02.07.26B) and subsequently gaining the permission of the Johns Hopkins CLUE Committee. The data were received in six different databases corresponding to the three epidemiological studies and two time periods. The data were manipulated so that the variables could be compared to each other, and then merged into one large database for each CLUE study. In total, the blood samples of 2,184 subjects were used.

3.2.5. Risk Factor Data

Additional participant information, such as age, race, gender, education level, marital status, and district average socioeconomic status (SES) as well as potential risk factors that have shown to be predictive of blood organochlorine levels such as body mass index (BMI), smoking status, drinking water source, and alcohol, fish, egg, dairy, vegetable and fruit consumption were obtained from the Johns Hopkins Training Center for Public Health Research (*45,63,69,79,155,165-167,189*). These data were abstracted from the CLUE I and II questionnaires, the 1975 private Washington County census, and the food frequency questionnaire. In addition, the spatial variables of direction and Euclidian distance from the Superfund site to the residence as well as urban/rural residence were created and added to the list of potential risk factors. These data were linked via subject ID to the biomarker data and combined into one database for analysis.

3.3. Statistics

The statistical modeling in this study involves regression analyses for spatially dependent data. The models are of the general form:

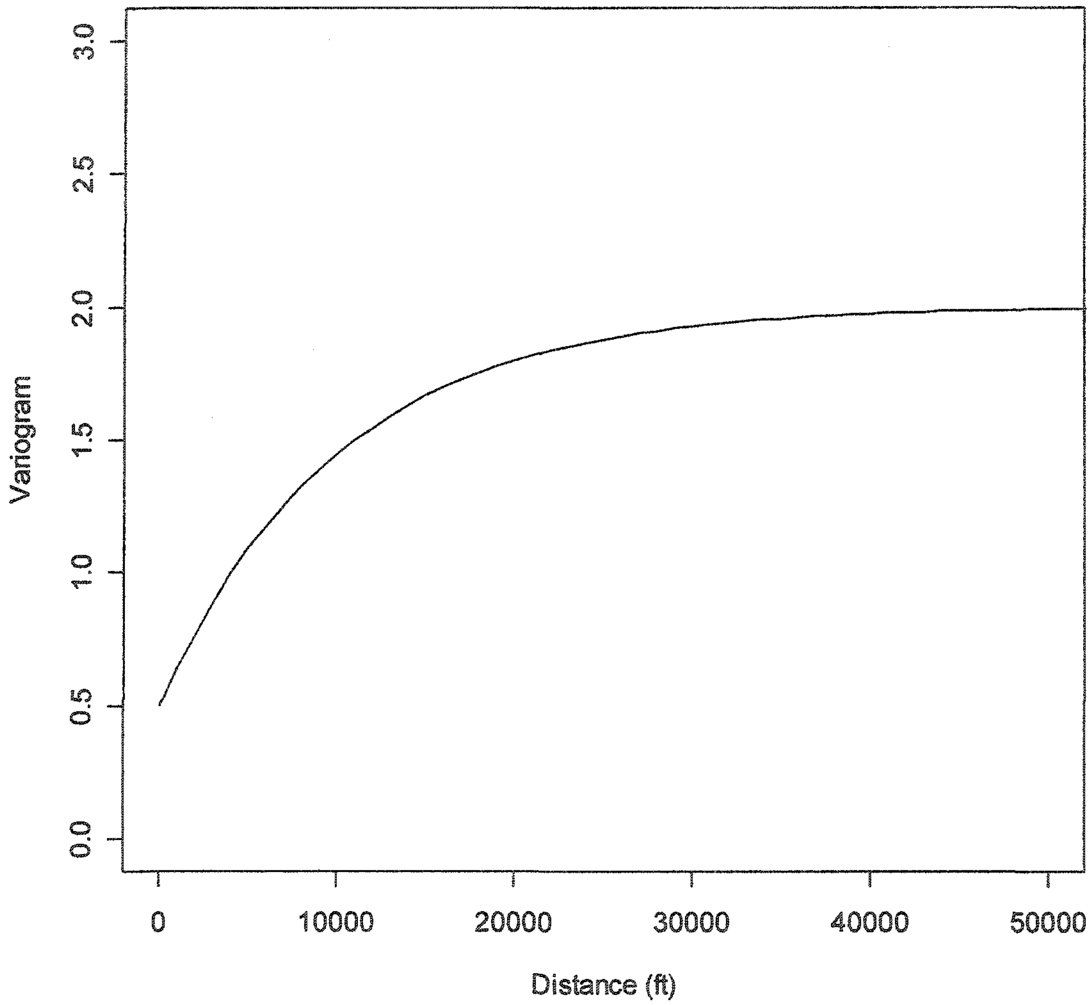
$$Y(\mathbf{s}) = \beta_0 + \beta_1 X_1(\mathbf{s}) + \ldots + \beta_n X_n(\mathbf{s}) + \varepsilon(\mathbf{s}), \qquad (3.1)$$

where '**s**' denotes spatial coordinates, Y(**s**) represents continuous outcomes at location **s**, $X_1(\mathbf{s})\ldots X_n(\mathbf{s})$'s risk factors (including possible interactions) indexed by location **s**, $\beta_1\ldots\beta_n$'s, their associated effects, and β_0 the baseline intercept. The residual error term, ε(s) is assumed normally distributed with a zero mean and constant variance. Model 3.1 resembles the common linear regression model with one major distinction; the error component, ε(**s**), is assumed to be spatially dependent (*190*). In the geostatistical literature, Model 3.1, with these specifications, is known as a universal kriging model commonly used for spatial prediction at unobserved or unmeasured locations (*23*). Various forms of Model 3.1 are used throughout the three manuscripts of this thesis for purposes related to spatial prediction, quantifying potential risk factors, and simulation.

The spatial dependence is accounted for in Model 3.1 by allowing the residual correlation between location pairs, denoted by ρ (ε($\mathbf{s}_i$), ε($\mathbf{s}_j$)), to be represented as a decreasing function of the distance between locations. There exists a pool of valid correlation functions that are routinely used in geostatistics, each parameterized to allow the data to help determine its complete structure (*23*). One such function known as the exponential correlation function, commonly used throughout the literature and in the second and third manuscripts, is given by $\rho^{d_{ij}}$, for $0<\rho<1$, with d_{ij} representing the distance between locations $\mathbf{s}_i$ and $\mathbf{s}_j$. Details regarding other correlation functions including the spherical correlation function used in the first manuscript can be found elsewhere (*23*).

The characterization of a spatial correlation function including parameter estimation is carried out in this research with a variogram (*23*). For instance, the variogram of the residual process in Model 3.1, is defined to be Var($\varepsilon(s_i)$, $\varepsilon(s_j)$), where Var denotes variance. Under certain regularity conditions, there is a one-to-one correspondence between variogram and correlation functions (*23*). An example exponential variogram (which would relate to the exponential correlation function mentioned above) is shown in Figure 3.2. The x-axis represents the distance between two data points. The y-axis represents the variance between two data values that are located at a specific distance from each other. The slowly rising and then leveling off pattern is indicative of spatial dependence since it shows that there is little variance between data values that are spatially close to one another, and the variance increases as the distance gets larger until the data values are no longer spatially dependent (variance becomes constant). A horizontal structure with no apparent pattern would suggest a lack of spatial dependence. Therefore, the variogram is commonly employed as a diagnostic check for assessing spatial dependence in data (*23*).

Figure 3.2: Example of variogram showing spatial dependence in data



A useful tool in practice and in this research is the following method of moments variogram estimator.

$$2\gamma(h) = \frac{1}{|N(h)|}\sum_{N(h)}\left(\varepsilon(s_i) - \varepsilon(s_j)\right)^2 \qquad (3.2)$$

where 2γ(h) denotes the variogram, γ(h), the semivariogram, h represents distance classes, N(h) the set of location pairs in each distance class, and |N(h)| the number of such pairs, with the sum (and hence, the average) taken over all the pairs of locations whose distance falls within the specified distance classes (*23*). This estimator is commonly employed as a diagnostic check for assessing spatial dependence in data (*23*). For analyses throughout this study, variogram function parameter estimation (equivalently correlation function estimation) for Model 3.1 was obtained via maximum likelihood based methods (*23,191*). However, the method of moments estimator (Model 3.2) was relied upon heavily as a diagnostic tool.

The second manuscript uses the universal kriging model (Model 3.1) to evaluate potential individual risk factors such as age, BMI, diet, smoking status, alcohol consumption, and drinking water source to determine their significance in influencing the levels of blood organochlorines in humans. A simple version of Model 3.1, known as ordinary kriging, is demonstrated in both the first and third manuscripts. This model assumes that the data do not depend on large scale trend. In other words, there are no known risk factors in the data. An example of an ordinary kriging model is as follows:

$$\mathrm{Y}(\mathbf{s}) = \beta_0 + \varepsilon(\mathbf{s}) \qquad (3.3)$$

where $\mathrm{Y}(\mathbf{s})$ is the outcome concentration value at location $\mathbf{s}$, β_0 is a regression coefficient representing a constant mean, and $\varepsilon(\mathbf{s})$ is a normally distributed spatially dependent error term representing residual variation. In the first manuscript, this model is used to predict levels of 1,1-dichloro-2,2-*bis*(*p*-chlorophenyl)ethylene (DDE) in soil surrounding the CCC Superfund site. The third manuscript also uses this ordinary kriging model in addition to conditional simulation to estimate the spatial variation in geocoding positional bias.

3.4. R Statistical Computing Environment

Specialized software packages exist to perform geostatistical analyses (*27-30*). The open source statistical computing environment R is used in this study. R is similar to the S and S-Plus statistical computing package (MathSoft), with main advantages being that it is free and continually expanded with contributed libraries for a wide range of specialized topics. R has an extensive range of built-in statistical techniques that can be used to perform standard statistical analyses from exploratory to more confirmatory model-based analyses. R is compatible on PC, Unix, Linux, and Macintosh platforms. In addition, R has several hundred official open source add-on libraries for other specialized techniques with approximately ten percent of these libraries containing tools to perform geographic/spatial analyses (*28*). Furthermore, R has powerful and unique graphical capabilities as is evidenced in this research. R has widespread usage in the academic community and increasing usage in industrial and government sectors. More detailed discussions about R and its contributed libraries can found elsewhere (*31*). Appendix C contains selected R code used for the statistical analyses performed in this research.

One contributed library in particular, geoR, was used extensively for performing all geostatistical analyses (*192,193*). The geoR library contains a comprehensive set of functions to perform most traditional based geostatistical methods as well as likelihood based and Bayesian methods. A more detailed discussion and tutorial about geoR can be found online (*31,193*).

MANUSCRIPT ONE

GEOSTATISTICS AND GIS: TOOLS FOR CHARACTERIZING ENVIRONMENTAL CHARACTERIZATION

4. Manuscript One: Geostatistics and GIS: Tools for Characterizing Environmental Contamination

4.1. Publication

This study was accepted for publication in the Journal of Medical Systems on January 6, 2004.

4.2. Abstract

Geostatistics is a set of statistical techniques used in the analysis of geo-referenced data that can be applied to environmental contamination and remediation studies. In this paper, the 1,1-dichloro-2,2-*bis*(*p*-chlorophenyl)ethylene (DDE) contamination at a Superfund site in western Maryland is evaluated. Concern about the site and its future clean up has triggered interest within the community because residential development surrounds the area. Spatial statistical methods, of which geostatistics is a subset, are becoming increasingly popular, in part due to the availability of geographic information system (GIS) software in a variety of application packages. In this article, the joint use of ArcGIS software and the R statistical computing environment are demonstrated as an approach for comprehensive geostatistical analyses. The spatial regression method, kriging, is used to provide predictions of DDE levels at unsampled locations both within the site and the surrounding areas where residential development is ongoing.

4.3. Introduction

Evaluating spatial relationships and geographic determinants of health is a growing area of environmental and public health research. Traditional environmental health studies have evaluated temporal changes in diseases and their determinants. More recently, however, there is a growing interest in evaluating both spatial and temporal patterns of health and their environmental determinants. As a result, geographic information systems (GIS) have become increasingly popular in environmental health applications and to environmental health practitioners (*12-19*). As GIS becomes more widespread, various applications of GIS have evolved. Articles in the literature provide various definitions and applications of GIS as a medical, environmental, and public health information system. In its basic form, GIS is a database system with the distinguishing feature that it deals with spatially (or geographically) referenced data. "Where" in addition to "what" that is measured or observed is important and thus recorded and stored in a GIS database. With location information linked to data values, GIS becomes a visual database providing comprehensive mapping capabilities as it's most basic and fundamental construct. However, the functionality of a GIS extends beyond producing maps. Some of these more specialized features are highlighted in this study with an application involving environmental contamination at a Superfund site.

The purpose of this paper is to present combined components of GIS and the R statistical computing environment as tools for performing geostatistical and environmental based analyses. Data from a western Maryland Superfund site are used to demonstrate different features of GIS and R as they are needed to follow through a geostatistical analysis characterizing potential contamination at the site. Background information on the Superfund site, geostatistics, and the R statistical computing environment are provided first.

4.4. Background

4.4.1. Description of the Site

In the early 1930's a large chemical company built a 19-acre facility in western Maryland for the production of fertilizers and pesticides, including blending 1,1,1-trichloro-2,2-bis(*p*-chlorophenyl)ethane (DDT) until 1968. DDT and its derivatives, 1,1-dichloro-2,2-*bis*(*p*-chlorophenyl)ethylene (DDE) and 1,1-dichloro-2,2-*bis*(*p*-chlorophenyl)ethane (DDD), are classified by the United States Environmental Protection Agency (EPA) to be Group B2, probable human carcinogens (*6*). These compounds have also been found to be endocrine disrupters and acute doses can severely affect the nervous system (*6,36*). In 1972, the use and production of DDT was banned by the United States. The chemical company ceased all fertilizer and pesticide operations at the site in 1984, however many original structures remain today.

Subsequent site investigations involving soil, water, air and fish sampling during the 1970's, 1980's and 1990's indicated the presence and migration of these pesticides and other organic and inorganic contaminants to off site areas, some of which are currently under residential development (*8-10,170,188*). Furthermore, the municipality in which the site is located has a population of approximately 60,000 residents. The EPA placed the site on the National Priority List (NPL) for cleanup in 1997, designating it a Superfund site (*11,194*). For demonstration purposes, soil sample data collected on and near the site in the 1990's analyzed using both GIS and statistical techniques are presented in this paper.

4.4.2. Geostatistics and R

The area of statistics known as geostatistics is ideally suited for characterizing the spatial distribution of environmental contamination and for evaluating potential spatial determinants. Geostatistics is a set of statistical techniques used in the analysis of spatially referenced data (*21-23*). Originally developed in the mining industry to predict recoverable ore reserves, the practice of geostatistics has found widespread use in environmental applications such as Superfund site characterization (*24-26,195,196*).

Geostatistics provides the ability to predict levels of contamination at unsampled locations based on known sampled values. Intuitively, predicting contamination levels at unsampled locations can be thought of as a weighted average of neighboring sampled values, with higher weight given to values closer to the prediction location. This can be accomplished formally using the following regression model,

$$Y(\mathbf{s}) = \mu(\mathbf{s}) + \varepsilon(\mathbf{s}), \tag{4.1}$$

where '$\mathbf{s}$' denotes spatial coordinates, $Y(\mathbf{s})$ represents contaminant values at location $\mathbf{s}$, $\mu(\mathbf{s})$ is the mean component, and $\varepsilon(\mathbf{s})$ corresponds to a zero mean normally distributed random error. Model 4.1 resembles the common linear regression model with one major distinction; the error component, $\varepsilon(\mathbf{s})$, is assumed to be spatially dependent. Levels of contamination at the site are likely to be more similar at samples closer together than samples further apart. This spatial dependence is accounted for in Model 4.1 by allowing the correlation between error terms to be structured as a decreasing function of the distance between sample locations.

There exists a pool of valid correlation functions that are routinely used in geostatistics, each parameterized to allow the data to help determine its complete structure. The function chosen for the correlation and, more importantly, the estimation of its parameters, comprise a crucial step in the geostatistical process. The completely specified correlation function determines the weighting scheme that yields optimal spatial predictions at unsampled locations, generating what are statistically known as best linear unbiased predictions (BLUPs). Once the correlation function has been determined, the BLUPs under Model 4.1 are generated via statistical least squares techniques (*23*).

A primary focus in geostatistics is thus specifying this correlation function. Traditionally, this has been accomplished with a closely related function known as the semivariogram (*23*). The spherical semivariogram function shown below is an example of a semivariogram commonly used in geostatistics (*23*),

$$\gamma(h) = \left\{ \begin{array}{ll} 0 & h = 0 \\ \tau^2 + \sigma^2 (1.5 * h / \varphi - 0.5 * (h / \varphi)^3) & 0 < h \leq \varphi \\ \tau^2 + \sigma^2 & h > \varphi. \end{array} \right\} \tag{4.2}$$

The parameter, φ, is the range, and corresponds to the distance (h) at which sample locations are no longer spatially correlated. Parameters τ^2 and σ^2, the nugget and partial sill, respectively, define the variance of the process. The nugget represents what is sometimes referred to as micro-scale variation, or variation between sample locations at a distance smaller than that observed in the sampling. By determining values for these parameters, the semivariogram is specified, and is easily transformed into a correlation structure used for generating BLUPs (*23*).

The traditional geostatistical approach first estimates the semivariogram based on the sampled data using the classic method of moments estimator, and then fits a semivariogram function through these estimates with a graphically based procedure such as weighted least squares (*23,197*). More recently, due to the subjectivity involved in this method, there has been a push towards a more model-based approach making use of statistical likelihood concepts for semivariogram parameter estimation (*191,198-200*).

Note when information in the form of covariates is available, the mean component, $\mu(\mathbf{s})$, in Model 4.1 may be parameterized as $\mu(\mathbf{s})= \beta_0 + \beta_1 X_1(\mathbf{s}) +\ldots+ \beta_n X_n(\mathbf{s})$, where the X's represent covariates indexed by location $\mathbf{s}$, and β's, their associated effects. In this case, Model 4.1 is known as a universal kriging model. If no auxiliary information is available, the mean component is assumed to be unknown and constant across the site, $\mu(\mathbf{s})=\mu$, yielding what is known as the ordinary kriging model. For those familiar with regression applications, this may seem unusual. However, prediction is often the primary goal of the kriging model, as opposed to estimation of covariate effects.

Specialized software packages exist to perform geostatistical analyses (*27-30*). The open source statistical computing environment R is used in this study. R is similar to the S and S-Plus statistical computing package (MathSoft), with main advantages being that it is free and continually expanded with contributed libraries for a wide range of specialized topics. R has an extensive range of built-in statistical techniques that can be used to perform standard statistical analyses from exploratory to more confirmatory model-based analyses. R is compatible on PC, Unix, Linux, and Macintosh platforms. In addition, R has over 300 official open source add-on libraries for other specialized techniques with several of these libraries containing tools to perform geographic/spatial analyses (*28*). Furthermore, R has powerful and unique graphical capabilities as is evidenced in this analysis. R has widespread usage in the academic community and increasing usage in industrial and government sectors. More detailed discussions about R and its contributed libraries are found elsewhere (*31*).

4.5. Methods

4.5.1. Environmental Data Source

Soil samples were collected at the western Maryland Superfund site and analyzed for DDT, DDE, and DDD among others at the request of the Maryland Department of the Environment (MDE). The results of these investigations were published in reports which were attained via the Freedom of Information Act (*8-10,170,188*). Within these reports, assayed values were presented in table form along with survey-type maps of the site that included drawings of buildings and other landmarks. Locations of sampled data were marked on these maps and could be cross-referenced to the assayed value using the corresponding unique sample identifier. However, actual geographic coordinates of each sample or building structure did not exist.

In soil, DDT usually breaks down into two derivate forms, DDE, a product of dehydrogenchlorination, and DDD, a product of dechlorination (*5,6*). For this study, 110 DDE surface samples (defined in this study as a depth of less than one foot) collected between 1992 and 1997 were used to characterize the spatial distribution of contaminants throughout the site. Multiple samples collected at the same location, time, and surface depth were averaged.

4.5.2. GIS

The DDE data were input into a Microsoft Excel database and imported into ArcGIS 8.3 Desktop (Environmental Systems Research Institute, Inc.). Geographic coordinates for each sample, needed to perform the geostatistical analyses, were determined by creating a georeferenced map of the site and surrounding area, digitizing all sample locations, and creating a geocoded database containing DDE concentrations by location.

Georeferencing is the process by which a map is made electronic, and geographically located on the globe (*201*). The process involves obtaining geographic coordinates of some landmark structures on the site. In this study, a hand held global positioning system (GPS) unit was used to record the geographic coordinates of the corners of each building and fence line during a site walk-through. Maps were scanned into ArcGIS and anchored, or georeferenced, using these landmark coordinates.

Geographic coordinates for the sample locations could not be obtained via a GPS since the ground did not contain any such markings to designate sample locations. Instead, sample locations were identified on the original site maps and made into points on the georeferenced ArcGIS site map with the landmark buildings and boundary fence lines in place. The computer mouse and ArcGIS editor tools were used to produce these points. This process is known as on-screen digitizing and resulted in an approximate set of geographic coordinates for the DDE sample locations (*201*).

In order to link the soil contaminant values with the soil sample locations, DDE concentrations were geocoded to the map using ArcGIS. Geocoding allows the user to link any piece of information, or attribute, to a geographical location which already exists in the GIS (*201*). In this case, levels of DDE found in soil are linked to the locations from which they were collected on the site layout map via a map identifier that was created when the sample location points were added to the map.

4.5.3. Geostatistics and R

The data, now spatially referenced, were analyzed using geostatistical techniques. The end result of the analysis is a map of the Superfund site characterizing predicted levels of DDE contamination based on the 1992-1997 soil samples.

The ordinary kriging regression model (Model 4.1) was used to predict levels of DDE at unsampled locations since no auxiliary information was available for the DDE sample data. The spherical semivariogram function (Model 4.2) was used for the correlation structure. Parameter estimates of the spherical semivariogram were obtained using the composite likelihood approach (*191*). The composite likelihood approach for semivariogram estimation has many of the advantages of a full statistical likelihood based approach with the robust and more positive properties of the traditional approach (*23,191*).

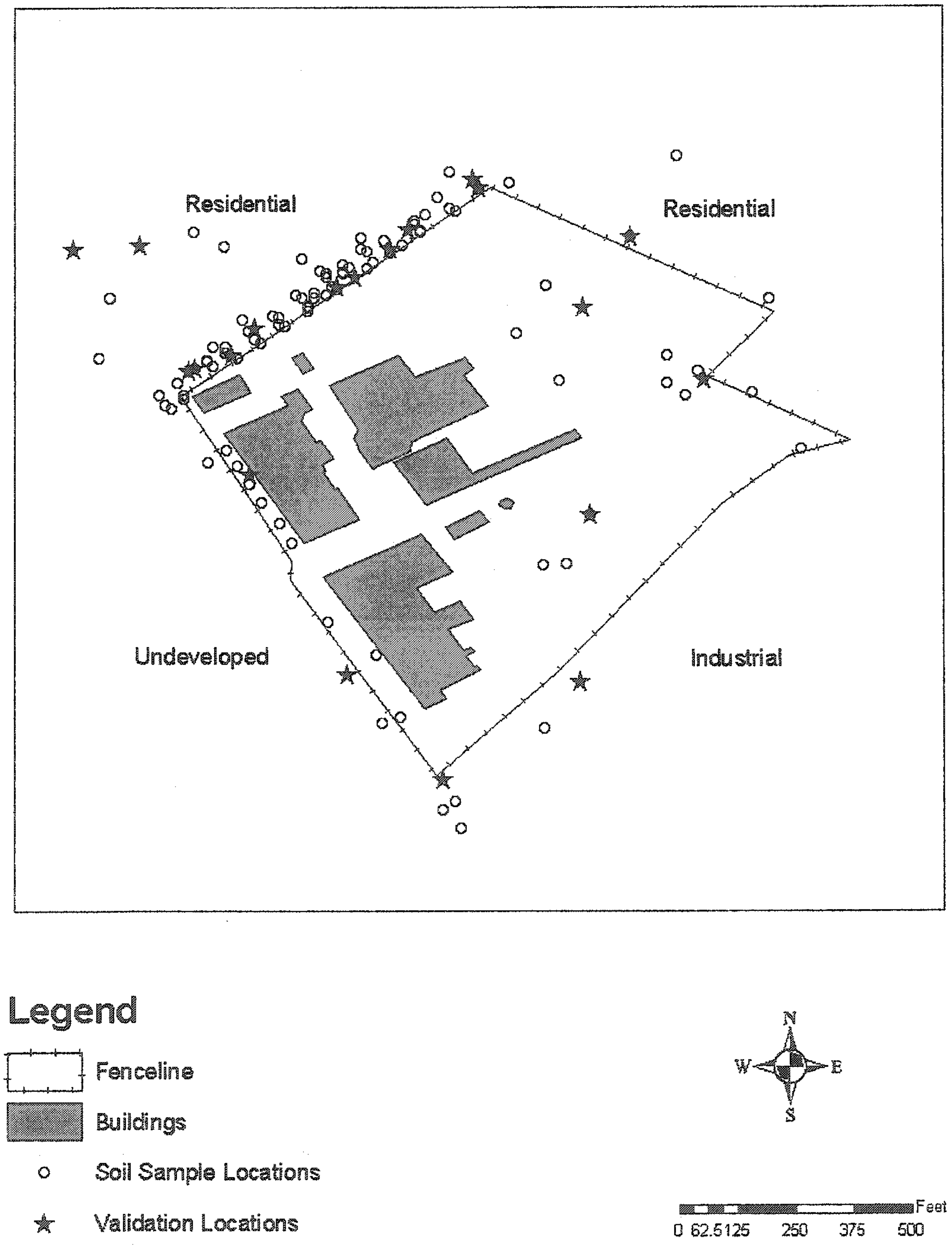
To assess predictive performance of the kriging model, five DDE samples were randomly chosen from each quartile of the 110 data values, and set aside for validation. These 20 data points did not contribute in the initial analysis. Instead, the remaining 90 values were analyzed and used to predict at these 20 validation locations. The root mean squared error between the true observed DDE values and the predicted DDE values was used to assess the model's ability to predict levels of DDE at unsampled locations. The geostatistical process was then repeated using the full set of 110 DDE samples to produce a map of predicted DDE contamination across the entire site.

All statistical computing and graphics (excluding ArcGIS maps) were performed within the open-source R statistical computing environment. The analysis conducted in this study used the contributed library geoR, a comprehensive set of functions to perform geostatistical based analyses (*192,193*). The geoR library contains written functions to perform most traditional based geostatistical methods as well as many likelihood based and Bayesian methods. A more detailed discussion and tutorial about geoR can be found online (*31,193*). The flexibility of R also allowed programming the composite likelihood method for semivariogram estimation.

4.6. Results

Georeferencing the original Superfund site map as well as digitizing and geocoding the soil sample locations yielded a fully functional GIS map (Figure 4.1). The map of the Superfund site displays buildings, boundaries, and the 1992-1997 DDE soil sample locations used for analysis. Buildings on the site included an old pesticide plant (which burned down), fertilizer plant, mixing building, warehouse, laboratory, maintenance shop, electrical shop, and boiler house. The site is bordered by railroad tracks and undeveloped land on the southwest side, a road and industrial property on the southeast border, and residential developments along the northwest and northeast fence lines.

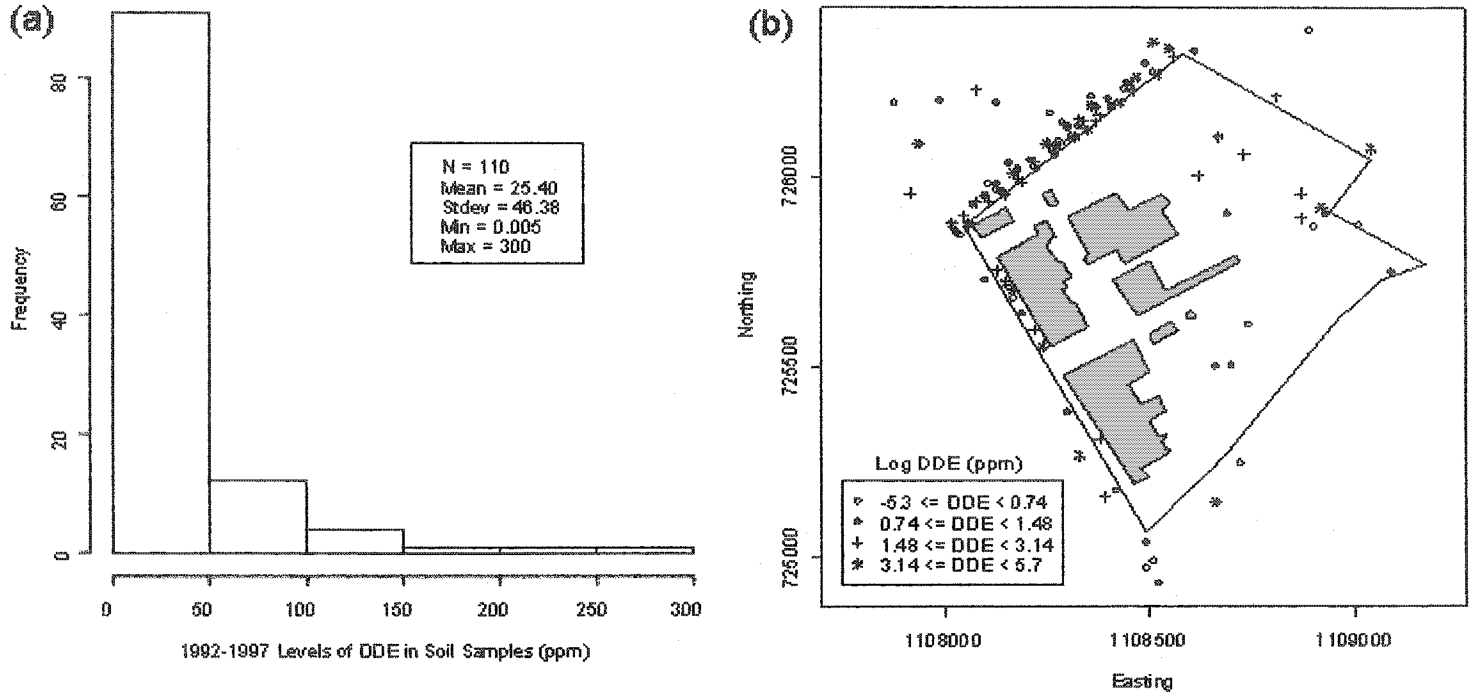
Figure 4.1: A GIS based map of the western Maryland Superfund site. Included are the 110 DDE soil sample locations from 1992-1997 indicating the 20 locations set aside for model validation



Soil samples collected throughout the site for various reasons characterize the extent of contamination (*8-10,170,188*). Intensive sampling was performed along the northwest fence line to determine whether the fence should be extended to contain any contamination that had migrated or spilled offsite. Single-family homes were being built in this region and continue to be built to this day. As a result of this sampling, the EPA mandated that the fence be extended by 20 feet to protect the health of those residents in 1996 (*188*).

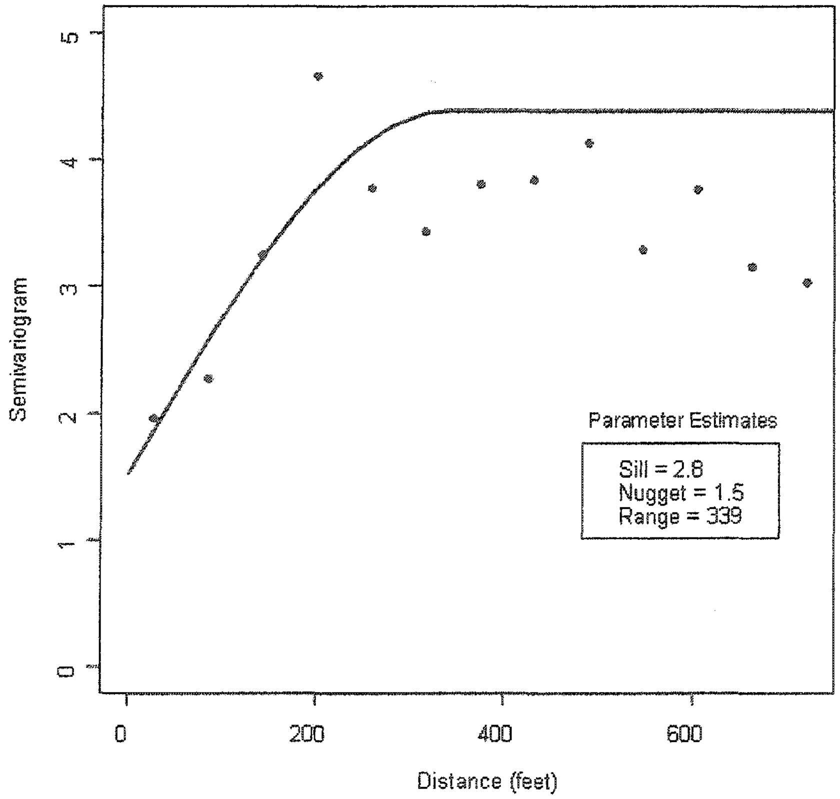
From 1992 to 1997, 110 soil samples were taken at the site and analyzed for DDE. The distribution of the DDE sample values were highly skewed to the right and ranged over several orders of magnitude (Figure 4.2a). Based on subsequent model validations, the DDE samples were log transformed for all analyses with final predictions transformed back to the original scale. Such an approach is common in geostatistics for environmental data with similar characteristics (*23,202-204*). An initial summary of the spatial distribution for the transformed DDE samples is shown in Figure 4.2b. The apparent clustering of similar values in this plot, with respect to the logged DDE quartiles, is evidence of spatial dependence. Also of interest is the close proximity between several extremely high and moderately low sample values. For example, on the original scale, levels of 0.5 parts per million (ppm) DDE were found in samples within 100 feet of samples with 28 ppm DDE.

Figure 4.2: Spatial and non-spatial exploratory analysis of the 110 DDE samples collected at the Superfund site, 1992-1997. (a) Histogram of the assayed DDE samples. (b) Spatial data posting based on quartiles for the log transformed DDE samples



Spatial dependence is more formally explored by estimating a semivariogram using the classic method of moments estimator, which is shown as 13 summary points in Figure 4.3. The rising and leveling-off pattern in the estimated semivariogram is an indication that the DDE sample values are spatially dependent. Lack of spatial dependence would be evidenced by a horizontal band of estimates showing no pattern. The composite likelihood estimated spherical semivariogram function is also shown in Figure 4.3 with a solid line. The range parameter was estimated to be approximately 339 feet beyond which DDE soils samples are assumed to be no longer spatially correlated.

Figure 4.3: Method of moments semivariogram estimates (points), composite likelihood fitted spherical semivariogram function (solid line), and corresponding parameter estimates for the log transformed DDE samples



The semivariogram estimates shown summarize spatial dependence for only a discrete set of distances. It is widely known that the methods of moments estimator is subjective and unreliable when based on a small number of spatial samples (*191,199*). The 110 DDE samples is not a large sample size in this regard (*200*). Therefore, the composite likelihood method was used for estimating the spherical semivariogram function because it is based on manipulations involving all 110 data samples, not just the summarized 13 estimates. For this reason, it is expected that the resulting estimated semivariogram function will not necessarily visually appear to fit these estimates well. Rather, the estimated semivariogram is used as an exploratory and diagnostic tool.

As an intermediate step in the analysis, the geostatistical process was validated using the 20 validation samples as described in the methods section. The spatial locations of this validation set are shown in Figure 4.1. Spatial dependence for the remaining 90 DDE values was estimated as outlined above and then used to predict at these validation locations. The predicted as well as measured DDE values are listed in Table 4.1. The overall root mean squared error in prediction (RMSEP) at these 20 validation locations was approximately 23.5 ppm. For validation locations in residential areas, the RMSEP was 16.9 ppm versus a RMSEP of 22.3 ppm for within site predictions.

Table 4.1: Measured Levels of DDE at the 20 Validation Locations, their Predicted Values, Prediction Errors*, and Area Designation According to Figure 4.1. Also listed are the Root Mean Square Error in Predictions (RMSEP) for the 20 Locations Overall and Stratified by Area

Area	Measured Value (ppm)	Predicted Value (ppm)	Prediction Error* (ppm)	Area	Measured Value (ppm)	Predicted Value (ppm)	Prediction Error* (ppm)
Site	1.7	8.2	6.5	Industrial	15.0	5.1	-9.9
Site	2.7	19.3	16.6	Industrial	92.0	153.1	61.1
Site	2.5	15.2	12.7	Residential	1.4	6.2	4.8
Site	3.1	44.1	41.0	Residential	0.6	21.7	21.1
Site	13.0	32.9	19.9	Residential	1.2	36.6	35.4
Site	8.1	54.9	46.8	Residential	1.8	17.5	15.7
Site	67.0	69.1	2.1	Residential	2.3	10.8	8.6
Site	66.6	77.2	10.5	Residential	3.8	11.0	7.2
Site	25.0	29.5	4.5	Residential	19.0	28.4	9.4
Site	47.4	37.9	-9.5	Residential	17.0	27.4	10.4

Overall RMSEP (ppm)	23.5
RMSEP within Site (ppm)	22.3
RMSEP outside of site in Industrial Property (ppm)	43.8
RMSEP outside of site in Residential property (ppm)	16.9

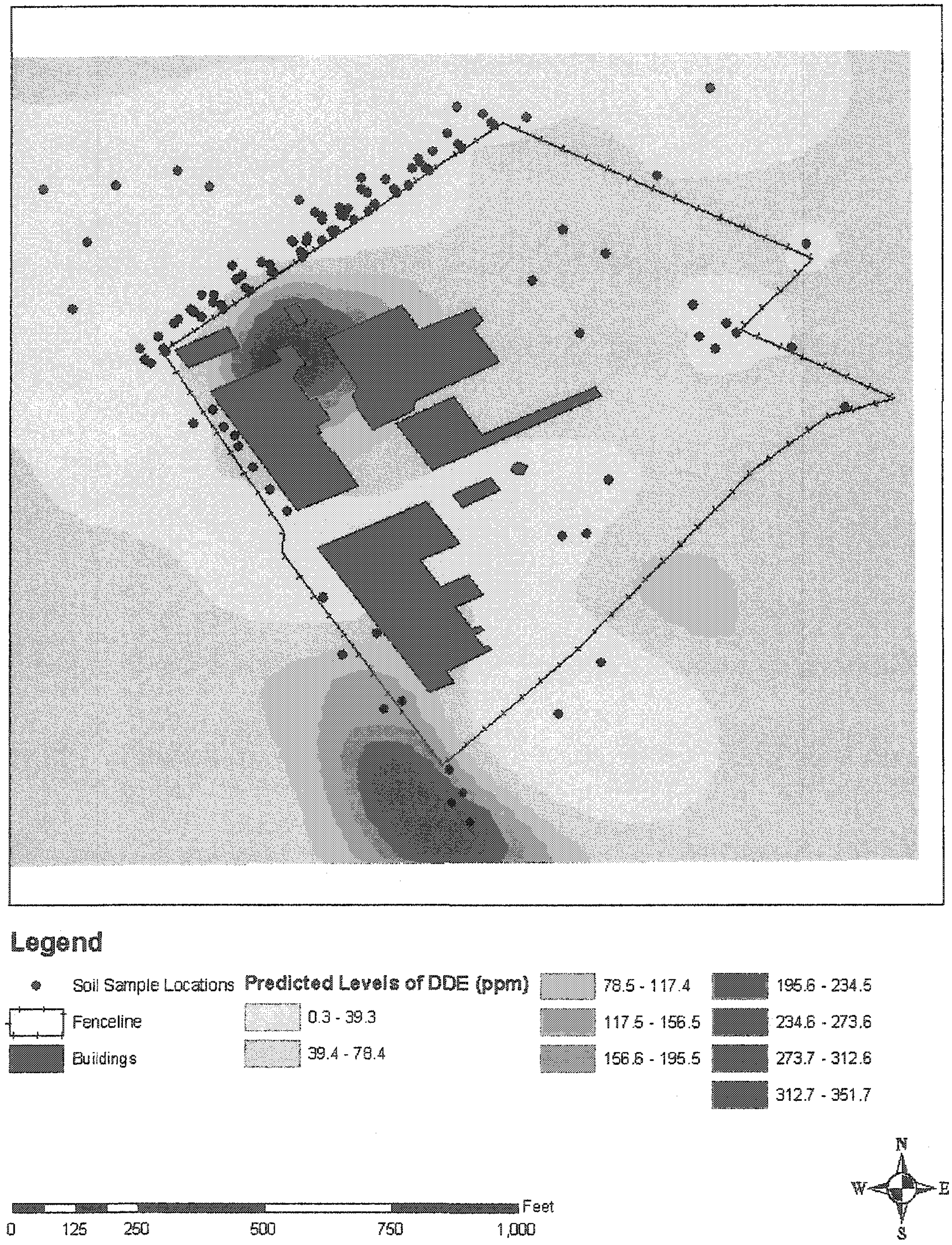
* Prediction Error = predicted value - measured value

Using all 110 DDE samples and the semivariogram function (Figure 4.3), kriged predictions (BLUPs) were generated at a grid of locations placed over the entire site and surrounding areas. Grid spacings were approximately 30 feet apart and there were a total of 2,500 prediction locations. The predictions were generated on the log scale and transformed back to the original scale (*23*). The predicted values were imported into ArcGIS and smoothed using inverse distance weighted interpolation to produce the map of DDE contamination (Figure 4.4).

The map indicates extremely high levels of predicted DDE in the northwest corner of the site and in the southwest area just beyond the fence line. Relatively low levels of DDE are predicted in the surrounding northwest and northeast regions of the site

(Figure 4.4). This is encouraging as these are areas of residential development. However, although these areas seem to have the lowest predicted values on the map, the predicted DDE levels in these regions are up to 39.3 ppm DDE. The EPA recommends cleaning up industrial sites to a contamination level of less than 10 ppm DDE (*205*). In addition, the EPA Region III risk based concentration (RBC) is 1.9 ppm for residential soil and 8.4 ppm for industrial soil based on standard exposure scenarios and a fixed level of risk (*206*).

Figure 4.4: Kriged map of predicted DDE levels for the western Maryland Superfund site, based on soil data collected between 1992-1997.



4.7. Discussion

In this paper, GIS and R, together were used to spatially analyze DDE contamination across a Superfund site in western Maryland. Via georeferencing, digitizing, and geocoding techniques, GIS provided the means to attain geographic coordinates of sample locations that otherwise were unknown, and required to spatially analyze the data. The statistical computing environment, R with the accompanying library, geoR, provide the necessary tools to carry out the geostatistical analysis that resulted in a map of predicted DDE levels across the site.

The analysis in this study is for demonstration purposes, and does not reflect a complete analysis of contamination at the site. A more complete analysis would include, for example, DDT and DDD, as well as other known contaminants. In addition, a map of prediction uncertainty for the DDE analysis would be of value to complement predictions and identify areas in need of further sampling. Additional samples are currently being collected by the EPA on and nearby the site to further determine the extent of contamination and identify the best remediation strategy.

A problem encountered in the analysis was the large variability in the sample data, especially that noted for samples close together. The most likely cause of the large variability in contaminant levels in the soil is spillage. Since the site blended DDT in the past, it is likely that there were several points of spillage. In fact, several historical reports mention observed signs of spills (*8-10,170*). The spillage of DDT could cause high levels of its derivative, DDE, to remain in these locations for long periods of time since DDE is not a mobile chemical in soil (*6*). Another smaller source of variability could be due the fact that the samples were taken over a period of five years. It is possible, although unlikely, that environmental degradation over time could have caused the levels of DDE in the soil to change. Furthermore, each year the samples were collected, they were analyzed by different laboratories. This source of variability is most likely small compared to the values observed in the soil.

Both GIS and R have capabilities beyond what were used in this study. For example, GIS can serve as a relational database by providing a common format to link specific data to additional geographical or environmental data collected for other purposes in the same geographic area. Each of these databases can be mapped, and layered atop of one another in a process called overlaying (*201*). For example, maps with property lines, proposed development, runoff patterns, and soil types could be overlayed with the DDE data used in this study. This is an avenue for further study.

Likewise, R has the capabilities of performing additional geostatistical techniques as demonstrated in the online geoR tutorial and related publications (*192,193,207*). With the contributed library, geoRglm, all methods for model-based statistics as described in Diggle *et al.* are available (*198,208*). In addition, there are several contributed libraries currently available in R devoted to geostatistics and other methods of spatial statistics. Together, R and GIS make optimal tools for data analysis in the environmental and public health fields.

4.8. Conclusion

This paper presents an evaluation of environmental contamination data that were analyzed using GIS and R to predict levels of contamination on and around a Superfund site in western Maryland. The methods and results of this study are important in the public health field because there are over 1,200 Superfund sites across the country that are contaminated with substances that adversely affect human health (*11*). Furthermore, Superfund sites are often located in urban areas surrounded by residences. Therefore, the health of people living near sites like these is dependant on the ability to accurately characterize and evaluate environmental contamination on and around sites like these. Spatial analyses, such as performed in this study using GIS and R, help to improve such characterization, thereby providing important methods in protecting the public's health.

4.9. Acknowledgements

This study was supported by the National Institute for Environmental Health Sciences (NIEHS) training grant (ES 07141), the Johns Hopkins NIEHS Center in Urban Environmental Health (P30 ES 03819) and the National Institute for Occupational Safety and Health Education and Research Center training grant (T42-CCT310419). In addition, the authors would like to thank Humberto Monsalvo, of the EPA for his helpful cooperation in obtaining data documents and guidance along site visits.

MANUSCRIPT TWO

THE USE OF SPATIAL INFORMATION IN DETERMINING POTENTIAL RISK FACTORS OF BLOOD ORGANOCHLORINE LEVELS IN A POPULATION LIVING NEAR A POTENTIAL SOURCE

5. Manuscript Two: The Use of Spatial Information in Determining Potential Risk Factors of Blood Organochlorine Levels in a Population Living Near a Potential Source

5.1. Abstract

The purpose of this paper is to evaluate residential location as a potential exposure determinant by investigating the geographic distribution characteristics of organochlorine levels in approximately 2,188 human blood samples in a community with a potential source in 1974 and 1989.

Potential individual risk factors such as age, race, gender, education, smoking status, body mass index, diet, socioeconomic status, drinking water source and potential spatial risk factors such as residential distance and direction from the potential source, and urban versus rural residence derived from geocoding address information in a geographic information system (GIS), are evaluated in generalized least squares regression models that further account for residual spatial variation.

The results of this study suggest that residential location may be an important exposure determinant of organochlorine levels in the blood of Washington County, Maryland, residents. A significant relationship was found between blood dieldrin levels and residential distance from the potential source adjusting for age, gender, education, smoking status and drinking water source and taking residual spatial variation into account.

This study demonstrates that evaluating potential spatial risk factors and taking into account spatial dependence in data models attempting to explain biomarkers of

exposure is important. Spatial information may enable the researcher to detect a potential exposure pattern that may not be seen with only non-spatial variables. Furthermore, ignoring spatial dependence in model residuals is not only incorrect, but it may weaken the model and more importantly, bias model findings by incorrectly overestimating the effect of model parameters.

5.2. Introduction

Geographic or spatial information has long been used to study the environmental contamination patterns of persistent organochlorine pollutants. The environmental fate, transport, and transformation of these compounds are known (*1*). Although they are ubiquitous, research involving the geographic distribution of these chemicals suggests that residing near a source may be a risk factor for higher than background exposure to organochlorines (*3,4*). These studies are limited to environmental measures, and the spatial distribution of residences for individuals with biomarkers of organochlorine exposure has not been investigated. Organochlorine compounds are known to bioaccumulate in the adipose tissue of humans, posing a threat to those exposed, but also allowing researchers the opportunity to estimate exposure with biomarkers. Therefore, the importance of spatial information may extend beyond environmental data to measures of human internal dose. Spatially evaluating biomarkers of exposure is a logical extension of spatial analysis research.

In 1974 and 1989, a Johns Hopkins Training Center for Public Health Research sponsored the collection of two large blood specimen banks as part of the Campaign Against Cancer and Stroke (CLUE I) and Campaign Against Cancer and Heart Disease

(CLUE II), respectively. In each campaign, approximately one third of the entire Washington County, Maryland population volunteered to participate (*131*). A subset of these blood samples were subsequently analyzed for organochlorines (*60,131,148*).

Hagerstown, a municipality located in Washington County contains an old chemical plant that is currently on the National Priority List (NPL) for cleanup. The site produced pesticides, such as 1,1,1-trichloro-2,2-*bis*(*p*-chlorophenyl)ethane (DDT), and fertilizers from the 1930's until the 1980's. Hence, it is a potential source of organochlorine exposure for residents of the surrounding community. Levels of organochlorine compounds in the environment suspected to be from the site have been previously characterized by the Maryland Department of the Environment (MDE) and the United States Environmental Protection Agency (EPA) and have been shown to have migrated off site (*8-10,170,209*).

The purpose of this paper is to evaluate residential location as a potential exposure determinant by investigating the spatial distribution characteristics of residences for individuals with measured blood levels of persistent organochlorine compounds in a community with a potential source. Potential risk factors based on geographic location are often overlooked, and further, predictors of blood organochlorine levels are determined by ordinary least squares regression techniques assuming independent residuals or the absence of residual spatial variation (*3,69,165-167,189,210-212*). This paper evaluates spatial and other known potential risk factors in linear regression models of blood organochlorine levels. Residual spatial variation from these regressions not accounted for by the model is used to adjust potential risk factor estimation and their standard errors, providing proper tests of significance.

5.3. Background

5.3.1. Geostatistics

The area of statistics known as geostatistics is ideally suited for characterizing the geographic distribution of environmental contamination and for evaluating potential spatial determinants (*21-23*). Originally developed in the mining industry to predict recoverable ore reserves, the practice of geostatistics has found widespread use in environmental and public health related applications (*23*).

Spatial relationships in data are often explored using estimated variograms (*23*). Variograms are cornerstone to the field of geostatistics, and used as a tool for characterizing spatial dependence (*23*). Positive spatial dependence exists when data points that are closer together in space are more similar than those that are further apart. Variograms are a function of distance and display the level and extent of spatial dependence over a range of spatial scales. More details regarding variograms and their applications are found in the background and research methods sections as well as in the first manuscript.

5.3.2. Organochlorines

Organochlorines are chlorinated aliphatic and aromatic hydrocarbons that are used extensively across the globe. Organochlorines are used as pesticides, insecticides, herbicides, fungicides, preservatives, and industrial fluids. The focus of this research is on DDT and polychlorinated biphenyls (PCBs) because although many organochlorine residues can be found in humans, and several are detectable in almost everyone, DDT and PCBs are thought to be the main contributors to the overall organochlorine body burden in humans (*4,36*). In addition, DDT represents a chemical that was extensively used at the Superfund site in western Maryland, which is located in the community that will be the focus of this investigation. On the other hand, PCBs represent ubiquitous organochlorines that were not made or often used at the Superfund site since it contained only one PCB-containing electrical substation, owned by the City of Hagerstown (*8*). Spatial patterns and the impact of spatial determinants will be compared and contrasted using these two chemicals.

Although DDT was banned in the United States in 1972, it is still used extensively in the developing parts of the world, and continues to have a widespread presence in the environment (*6,40,43,45*). DDT and its derivatives, 1,1-dichloro-2,2-*bis*(*p*-chlorophenyl)ethylene (DDE) and 1,1-dichloro-2,2-*bis*(*p*-chlorophenyl)ethane (DDD), are classified by the EPA as Group B2, probable human carcinogens and as possible human carcinogens (Group 2B) by the International Agency for Research on Cancer (IARC) (*5,6,36,41,60,129-131*). In addition, these compounds have also been found to be endocrine disrupters and acute doses can severely affect the nervous system (*6,36*). PCBs have also been linked to several forms of adverse health effects, and have been classified as probable carcinogens by the EPA, IARC, the National Toxicology Program (NTP), and the National Institute for Occupational Safety and Health (NIOSH) (*5,106*).

5.3.3. Superfund Site

In the early 1930's a large chemical company built a 19-acre facility in Hagerstown, Maryland for the production of fertilizers and formulation of pesticides, including blending DDT until 1968. In the 1970's and 1980's several dumpsites were uncovered in the grounds surrounding the site. These dump areas contained naphthalene, volatile organic compounds, and pesticides as well as unconfirmed but suspicious materials. Site investigations involving soil, water, air and fish sampling during the 1970's, 1980's and 1990's indicated the presence and migration of organochlorine pesticides, such as DDT, DDE, DDD, hexachlorobenzene (HCB), aldrin, chlordane, dieldrin, and methoxychlor and other organic and inorganic contaminants to off site areas, some of which are residential (*8-10,170,188*). The site became inactive in 1984 (*170*). The EPA placed the site on the National Priority List (NPL) for cleanup in 1997, designating it a Superfund site (*11,194*). A more detailed description of the site can be found in the first manuscript.

5.3.4. Blood Collection

In 1962, the Johns Hopkins Training Center for Public Health Research was founded in western Maryland (*164*). Among its many accomplishments, the Center sponsored the collection of two large blood specimen banks as part of the CLUE I and CLUE II campaigns. Over 20,000 adult residents volunteered for CLUE I from August through November of 1974, approximately six years after the Superfund site stopped making pesticides, and ten years before it became inactive. These participants represented roughly one third of the County population. In May through October of 1989, the CLUE II campaign recruited approximately one third of the population to participate as well. An estimated 25 percent of the CLUE I participants also participated in CLUE II (*131*).

After signing a general consent form, volunteers donated their blood to be used for research purposes. A subset of these samples were analyzed for organochlorines, and used as biomarkers for organochlorine exposure in studies by Helzlsouer *et al.*, Rothman *et al.* and Cantor *et al* (*60,131,148*). All the blood samples were analyzed for DDT, DDE, and 28 PCB congeners. Less than half of the 2,188 blood samples were analyzed for additional organochlorines such as Aldrin, dieldrin (CLUE I only), chlordane, HCB, hexachlorocyclohexane (HCH), endrin, (CLUE I only), oxychlordane, *trans*-nonachlor, heptachlor, heptachlor epoxide, and mirex (*60,131*). Details concerning the blood collection, storage, and analytical methods are found in the background section and published elsewhere (*60,131,148*).

5.4. Methods

5.4.1. Data

A total of 1,389 samples were analyzed for organochlorines from the CLUE I initiative, and 795 samples were analyzed from the CLUE II recruitment campaign. Blood sample organochlorine levels from six subjects whose blood was analyzed twice due to their participation in two of the three cancer studies mentioned previously were averaged for analysis purposes.

Street addresses and zip codes of the study participants were collected as part of the CLUE campaigns. ArcGIS (Environmental Systems Research Institute, Inc.) software was used to geocode the addresses providing a corresponding set of longitude and latitude coordinates. The geocoding process employed several base maps (*177,213*). StreetMap 2000 (Environmental Systems Research Institute, Inc.), the United States

Census Bureau 2000 Topologically Integrated Geographic Encoding and Referencing (TIGER) street maps, and the Delorme Street Atlas software (Delorme) were all used in an effort to successfully geocode a large percentage of addresses. Addresses that could not be geocoded consisted mainly of post office boxes and rural routes, and were removed from analysis. In situations where several study participants had the same residential address, such as in an apartment building, the geocoded coordinates were altered slightly by adding one foot increments to each the longitude and latitude coordinates, ensuring all participants had a unique location and no two participants were grouped into one when mapping. The Superfund site address was also geocoded. Details regarding this geocoding process are found in the research methods section.

In addition to geocoded address information, participant demographic information, such as age, race, gender, education level, marital status, and district average socioeconomic status (SES) (CLUE I only) as well as variables that have shown to be predictive of blood organochlorine levels such as body mass index (BMI) (CLUE II only), smoking status (current smoker at the time of blood draw), drinking water source (municipal/well/spring) (CLUE I only), and alcohol, fish, egg, dairy, vegetable and fruit consumption (CLUE II only) were obtained *(45,63,69,79,155,165-167,189)*. These data were abstracted from CLUE-based questionnaires and a private Washington County census. The spatial variables of direction and Euclidian distance from the Superfund site to the residence as well as urban/rural residence were created and added to the list of potential risk factors. Direction was determined based on residential location in one of 24 directional bins (every 15 degrees) from the Superfund site. Urban residence was defined in this paper to be living within 1.5 miles of the center of Hagerstown City.

Since DDT breaks down to DDE in the blood, DDE levels were chosen to represent DDT exposure (*4*). A total of nineteen PCB congeners were evaluated. However, only congeners 105, 118, 146, 153, 156, 170 and 180 were used for analysis. These congeners represent those that are more commonly found in the general population, and had the largest number of samples available for analysis (*89,214*). Total PCBs refer to the addition of all the congeners listed above. Congener 156 was omitted from the CLUE II (1989) analysis since only approximately half the CLUE II blood samples were analyzed for this congener. Dieldrin was also chosen to represent an additional compound that was used at the Superfund site.

In order to account for the effect of not fasting before blood draw, lipid adjusted values of all organochlorines were calculated according to the methods of Philips *et al.*, taking into account cholesterol and triglycerides (*169*). Some analytical labs reported levels below the official limits of detection (LOD), which were used since they are considered more valid than an interpolated value (*131,148*). The LOD divided by the square root of two was used for the unreported values below detection (less than five percent of the samples were below the LOD) (*215*).

5.4.2. Statistical Analysis

Summary statistics were generated for each compound considered, and stratified for CLUE I (1974) and CLUE II (1989). Levels of these organochlorines in participant's blood were mapped using their geocoded coordinates to study the spatial distribution of the levels of these organochlorines and their possible relationship to the Superfund site. Spatial structure in the levels of organochlorines was further explored using estimated variograms (*23*). Positive spatial dependence was suspected with the organochlorine blood data. Variograms were also estimated for regression residuals as a diagnostic check on the independence assumption inherent in ordinary least squares regression (OLS) inference.

Multivariate linear regression was used to develop models that best describe the blood levels of each organochlorine, both lipid adjusted and unadjusted, in both 1974 (CLUE I) and 1989 (CLUE II). These models are of the form:

$$Y(\mathbf{s}) = \beta_0 + \beta_1 X_1(\mathbf{s}) + \ldots + \beta_n X_n(\mathbf{s}) + \varepsilon(\mathbf{s}), \tag{5.1}$$

where '$\mathbf{s}$' denotes spatial coordinates, $Y(\mathbf{s})$ represents blood organochlorine levels of participants residing at location $\mathbf{s}$, $X_1(\mathbf{s})\ldots X_n(\mathbf{s})$'s risk factors (including possible interactions) indexed by location $\mathbf{s}$, $\beta_1\ldots\beta_n$'s, their associated effects, and β_0 the baseline intercept. The residual error term, $\varepsilon(\mathbf{s})$ was assumed normally distributed with a zero mean and constant variance. To further account for possible residual spatial variation, residuals were allowed to be spatially dependent by parameterizing their correlation as a decreasing function of the distance between their locations. In the geostatistical literature, Model 5.1, with these specifications, is known as a universal kriging model commonly used for spatial prediction at unobserved or unmeasured locations (*23*). However, the goal in this paper is inferential with a primary focus in selecting and quantifying potential risk factors that best predict blood organochlorine levels.

The model selection process began by running all possible models derived from each combination of potential risk factors considered as regression main effects. All potential risk factors were checked for colinearity, and those found to be correlated with one another were evaluated separately in the models to determine which were the best predictors. The fraction of variance explained by the model adjusted for the number of explanatory variables (adjusted R^2) as well as Akaike's Information Criterion (AIC) were used to rank model performance (*216*). The top ten to twenty performing models were then further investigated for significant interactions among the included potential risk factors. The final models chosen depended on model parsimony and scientifically meaningful interpretations. All regression inference at this step was based on OLS regression that assumes uncorrelated or independent residuals, which is possibly not a valid assumption. OLS was used because the objective was to first arrive at a manageable set of plausible final models to ultimately investigate and correct for residual dependence. OLS methods for estimating regression parameters in models with dependent residuals would lead to spurious significant inclusion of potential risk factors, which would then be re-evaluated adjusting for residual spatial dependence (*217,218*).

The set of final models arrived at from above were adjusted for possible residual spatial dependence using a two stage approach (*23,219*). Variogram functions were estimated for the OLS residuals and used to specify the residual correlation structure. The residual correlation structure was then used to re-estimate model effects using generalized least squares (GLS) methods (*190*). To reduce bias from using an estimated correlation structure based on the OLS model, another variogram was produced based on the GLS residuals, and a new correlation structure was estimated. This structure was then used to further update the GLS estimated effects, completing the two stage approach.

The exponential variogram function, routinely applied in spatial statistics, was used to characterize spatial dependence for all model residuals and estimated using maximum likelihood techniques (*23*). All statistical analysis was done using the open source R statistical computing environment with the contributed package, geoR, for spatial statistical operations (*31,192*).

5.5. Results

Demographic information regarding the study population is found in Table 5.1. The mean age of the CLUE I participants was 53 years old. The participants were 58 percent male and 19 percent were current smokers. The CLUE II population consisted of 68 percent men and 11 percent current smokers with a mean age of 63 years. Both populations were 98 percent white (two percent African American) with a mean education level of between 11 and 12 years. Because there were so few African Americans, the results are limited to only the white population with a brief explanation of how the results differ when African Americans are added to the analysis.

Table 5.1: Summary of Demographic Information for CLUE I (1974) and CLUE II (1989) Participants

	CLUE I: N=1374	CLUE II: N=787
Mean Age (years)	53.1	64.06
Whites	98.3%	98.1%
Mean Education (years)	11.24	11.92
Mean BMI	NA	26.35
Current Smokers	25.9%	11.1%
Men	57.8%	68.6%
City Water Drinkers	66.7%	NA
Spring Water Drinkers	0.7%	NA
Well Water Drinkers	11.8%	NA
Urban Residents	35.8%	25.9%
Mean Distance to Site (miles)	3.97	4.5

NA: Not Applicable to Dataset

The blood data were separated into two datasets corresponding to CLUE I (1974) and CLUE II (1989) since the number of samples as well as the available information on potential risk factors were not the same for each group. Table 5.2 summarizes the blood DDE, total PCB and dieldrin levels in the participants of each study. The mean level of DDE in the blood of Washington County residents adjusted for lipid content decreased from 1974 (3,024 ng/g) to 1989 (1512 ng/g) by approximately 50%. This is in accordance with a trend seen throughout the literature indicating that body burdens of persistent organochlorines have been declining since their use has decreased (*37,211,212*). However, even if the lower 95 percent confidence level of CLUE II mean DDE was to decrease another 50 percent, the mean levels of DDE in this population are well above the Second National Health and Nutrition Examination Survey (NHANES II) reported national background level of 297 ng/g in 1999 to 2000 (*4*).

Table 5.2: Summary of Blood Organochlorine Data for both CLUE I (1974) and CLUE II (1989) Datasets

	CLUE I				CLUE II			
Organochlorine	Number of samples	Range	Mean ± Standard Deviation	95% Confidence Interval	Number of samples	Range	Mean ± Standard Deviation	95% Confidence Interval
p,p DDE								
lipid adjusted (ng/g)	1274	36.4-25205	3023.5 ± 66.3	(2893.5, 3153.5)	659	38.0-12180	1512.0 ± 47.0	(1420.0, 1604.0)
unadjusted (ng/ml)	1374	0.19-185.9	20.0 ± 0.45	(19.1, 20.8)	786	0.27-61.9	8.3 ± 0.25	(7.8, 8.8)
Total of PCB congeners: 105,118,146,153,156,170,180								
lipid adjusted (ng/g)	1046	129.4-9355	771.5 ± 22.0	(728.4, 814.7)	NA	NA	NA	NA
unadjusted (ng/ml)	1049	0.73-63.9	5.0 ± 0.15	(4.7, 5.2)	NA	NA	NA	NA
Total of PCB congeners: 105,118,146,153,170,180								
lipid adjusted (ng/g)	NA	NA	NA	NA	617	47.6-3154	517.4 ± 15.7	(486.7, 548.1)
unadjusted (ng/ml)	NA	NA	NA	NA	658	0.35-25.5	2.7 ± 0.08	(2.6, 2.9)
Dieldrin								
lipid adjusted (ng/g)	685	1.7-825	113.07 ± 3.05	(107.1, 119.1)	NA	NA	NA	NA
unadjusted (ng/ml)	688	0.01-5.6	0.72 ± 0.02	(0.7, 0.8)	NA	NA	NA	NA

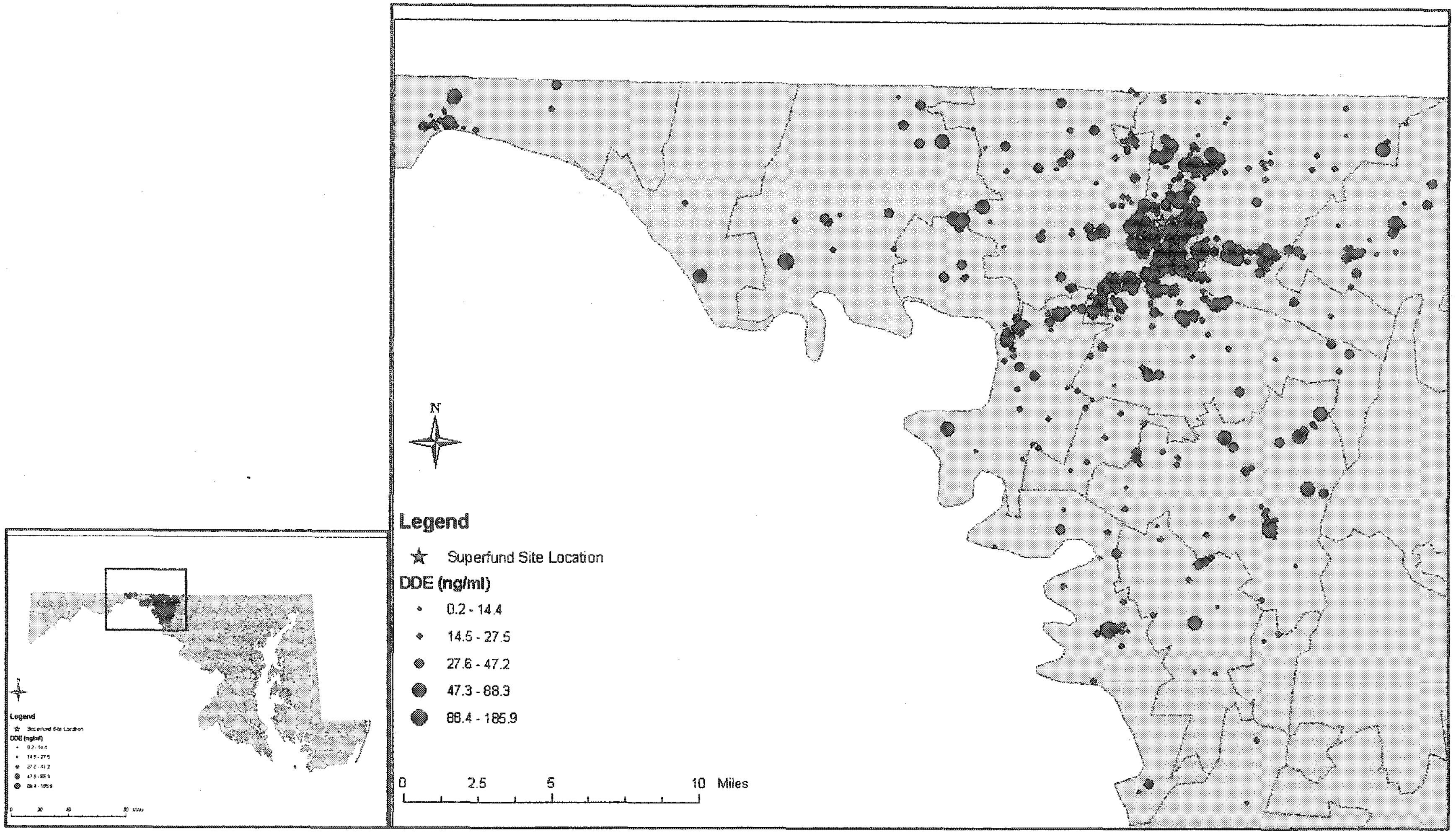
NA: Not evaluated in this study

In total, approximately 96 percent of the CLUE I and 95 percent of the CLUE II addresses were able to be geocoded. The CLUE I residential locations and levels of blood DDE are shown in Figure 5.1. The relative level of blood DDE in the Washington County residents are represented by the size of the point at their residential location. Additional maps can be found in Appendix A. Aside from clustering of residences in accordance with population density, spatial patterns were not apparent.

The addresses that were not geocoded (50 from CLUE I and 31 from CLUE II) consisted mainly of rural routes and post office boxes that the basemap was unable to locate (*177,178*). They showed no distinct patterns with respect to potential risk factors such as age, gender, smoking status, and were from various zip codes, uniformly dispersed across the County. There exclusion from the analysis, therefore, is not expected to introduce bias.

From the exhaustive search, plausible regression models were chosen for each organochlorine. Spatial dependence was found in the residuals of all organochlorines in this step as diagnosed by their estimated residual variograms. Parameter estimates and tests of significance were adjusted for this residual spatial dependence using the two stage GLS based approach outlined previously.

Figure 5.1: Map showing the geographic distribution of blood DDE levels in Washington County, Maryland



The impact each potential risk factor had on the blood organochlorine level alone was also measured using univariate GLS two stage regression. Tables 5.3 and 5.4 describe the results of the crude (univariate) and adjusted (multivariate) models used to describe the blood levels of DDE, total PCBs, and dieldrin in the Washington County residents in 1974 and 1989, respectively. Although all explanatory variables mentioned in the methods section were evaluated in the regression analysis, only those potential risk factors that were found to be predictive of blood organochlorines are presented in these Tables.

The results of the GLS regression modeling of lipid adjusted and unadjusted blood DDE levels including estimated parameter coefficients and corresponding p-values are found in Table 5.3. In 1974, age, gender, smoking status, education, drinking water source and distance to the Superfund site were found to improve the overall fit of the model of blood DDE levels. Women, non-smokers, and city water drinkers are found to have statistically significantly less DDE in there blood than men, smokers and those that drink spring or well water, respectively, when all other potential risk factors are equal. DDE levels are also found to increase significantly with age. No statistically significant association was found between the level of DDE in the blood and distance of the residence from the Superfund site as was suspected.

Table 5.3: Results of Generalized Least Squares Regression Determining the Effects of Covariates on Blood Organochlorine Levels, 1974: Parameter Coefficients (p-values)

Covariate	Compounds Used at CCC site						Compounds not used at CCC site			
	DDE (ng/ml)		DDE (lipid adjusted) (ng/g)		Dieldrin (lipid adjusted) (ng/g)		Total PCBs (ng/ml)		Total PCBs (lipid adjusted) (ng/g)	
	Crude	Adjusted	Crude	Adjusted	Crude	Adjusted	Crude	Adjusted	Crude	Adjusted
Distance from CCC site to Residence (miles)	0.14 (0.5)	0.12 (0.5)	14.1 (0.6)	9.1 (0.6)	-0.75 (0.32)	-1.6 (0.042)	NA	NA	NA	NA
Urban vs. Rural residence (rural vs. Hagerstown)	NA	NA	NA	NA	NA	NA	0.71 (0.013)	0.37 (0.2)	132 (0.002)	66.9 (0.1)
Age (years)	0.17 (<0.001)	0.12 (<0.001)	15.7 (0.01)	10.7 (0.072)	0.22 (0.4)	-0.19 (0.5)	0.02 (0.08)	0.01 (0.3)	0.57 (0.7)	-0.43 (0.8)
Gender (male vs. female)	11.2 (<0.001)	10.8 (<0.001)	1735.7 (<0.001)	1695 (<0.001)	16.8 (0.06)	14.6 (0.09)	4.3 (<0.001)	4.2 (<0.001)	641.4 (<0.001)	635 (<0.001)
Education (years)	-0.23 (0.2)	-0.12 (0.4)	-1.4 (0.9)	3.8 (0.9)	-1.5 (0.28)	-1.8 (0.2)	-0.05 (0.3)	-0.06 (0.11)	-2.4 (0.7)	-5.3 (0.4)
Smoking Status (non-smoker vs. smoker)	-2.3 (0.023)	-2.4 (0.013)	-291.1 (0.06)	-296.4 (0.042)	19.3 (0.01)	21.8 (0.004)	-0.37 (0.2)	-0.25 (0.3)	-60.5 (0.2)	-36.7 (0.3)
Drinking water Source (spring vs. municipal)	16.9 (<0.001)	15.4 (<0.001)	2089.5 (0.005)	1918.8 (0.004)	221.0 (<0.001)	221.2 (<0.001)	0.74 (0.6)	-0.17 (0.9)	26.7 (0.9)	-83.4 (0.6)
Drinking water Source (well vs. municipal)	3.0 (0.039)	2.5 (0.065)	406.2 (0.065)	362.6 (0.08)	3.4 (0.7)	9.4 (0.4)	0.88 (0.028)	0.27 (0.4)	140.4 (0.025)	48.9 (0.4)
Adjusted R squared		0.209		0.243		0.3376		0.428		0.36
Degrees of Freedom		1100		1030		588		861		859

NA: Not in final model

Table 5.4: Results of Generalized Least Squares Regression Determining the Effects of Covariates on Blood Organochlorine Levels, 1989: Parameter Coefficients (p-values)

Covariate	DDE (ng/ml)		DDE (lipid adjusted) (ng/g)		Total PCBs (ng/ml)		Total PCBs (lipid adjusted) (ng/g)	
	Crude	Adjusted	Crude	Adjusted	Crude	Adjusted	Crude	Adjusted
Urban vs. Rural residence (rural vs. Hagerstown)	NA	NA	NA	NA	NA	NA	-31.1 (0.6)	-13.5 (0.8)
Age (years)	0.04 (0.3)	0.1 (0.022)	10.2 (0.2)	14.8 (0.08)	0.04 (<0.001)	0.02 (0.002)	9.2 (0.001)	2.1 (0.4)
Gender (male vs. female)	-2.4 (0.004)	-3.1 (<0.001)	-81.5 (0.6)	-200.8 (0.2)	1.7 (<0.001)	1.6 (<0.001)	397.7 (<0.001)	374 (<0.001)
Education (years)	-0.35 (0.012)	-0.21 (0.11)	-62.9 (0.015)	-38.3 (0.2)	NA	NA	-19.5 (0.017)	-15.6 (0.038)
BMI	0.24 (0.013)	0.31 (0.001)	35.4 (0.04)	41.5 (0.015)	0.05 (0.014)	0.03 (0.07)	NA	NA
Smoking Status (non-smoker vs. smoker)	-3.2 (0.015)	-2.6 (0.045)	-597.5 (0.009)	-477.5 (0.044)	NA	NA	NA	NA
Often Dairy Consumption (everyday or more) vs. none	-0.6 (0.5)	-1.6 (0.3)	-38.4 (0.8)	-118.7 (0.6)	NA	NA	68.1 (0.2)	-16.8 (0.8)
Often Vegetable Consumption (everyday or more) vs. less than once a day	-4.1 (0.026)	-3.2 (0.08)	-1048 (0.002)	-836.2 (0.015)	NA	NA	NA	NA
Adjusted R squared		0.414		0.442		0.45		0.793
Degrees of Freedom		277		266		643		266

NA: Not in final model

The results of the GLS regression modeling of lipid adjusted blood dieldrin levels including estimated parameter coefficients and corresponding p-values are found in Table 5.3. After adjusting for age, gender, smoking status, education, and drinking water source, a statistically significant negative association was found between dieldrin levels in blood and the residential distance from the Superfund site. It is interesting to note that the only significant predictors of blood dieldrin levels were smoking status and drinking spring water versus city water. Furthermore, smokers tended to have significantly less dieldrin in their blood than non-smokers. Nonetheless, the results of this dieldrin model suggest that those that live closer to the site have higher levels of dieldrin in their blood than those that live further away.

As expected, no relationships between distance to the Superfund site and blood levels of total PCBs were found in the 1974 data as shown in Table 5.3. The spatial risk factor, urban versus rural residence, was significantly predictive of lipid unadjusted total PCB levels in blood when adjusting for age, gender, education, smoking status, and drinking water source. Blood levels of total PCBs in participants living within 1.5 miles of the center of Hagerstown are lower than those living outside of Hagerstown holding age, gender, education, smoking status, and drinking water source constant. The association is not significant for lipid unadjusted blood total PCB levels. In addition, while adjusting for other explanatory variables, men, smokers, and well water drinkers had higher blood PCB levels than women, non-smokers, and those that drink city water, respectively. However, only the association with gender is statistically significant. Finally, a positive association with age, and a negative association with years of education and blood PCBs were found, although neither of these relationships is statistically significant.

A summary of CLUE II results is presented in Table 5.4. With the exception of lipid adjusted total PCB's, the potential spatial risk factors (residential distance from Superfund site and urban/rural residence) were not found to be explanatory of blood organochlorine levels in 1989. Non-spatial risk factors, age, gender, BMI, education, smoking status, and vegetable and dairy consumption were found to improve the overall fit of the blood DDE model. Of these variables, only age and BMI were found to be statistically significantly positively associated with blood DDE levels. The relationships are slightly weaker when adjusted for the lipid content in the blood, however. Furthermore, men and non-smokers had lower blood DDE levels than women and smokers, respectively, adjusting for the other covariates. The association with gender was only statistically significant for the lipid unadjusted values of blood DDE. The relationship between smoking and blood DDE was statistically significant in both models.

In 1989, age, gender, and BMI were the only covariates found to improve the overall fit of the model of lipid unadjusted total PCBs in blood. Age was found to be significantly positively associated with blood PCBs. Education, dairy consumption and urban vs. rural residence in addition to age and gender were found to be potential risk factors for lipid adjusted total PCBs while only gender and education had statistically significant associations.

Results of this study indicate that the impact of spatial information varies within and between each time period. For example, the results show a statistically significant association between residential distance from the Superfund site and blood dieldrin levels in 1974. Weaker associations were found between residential distance from the site and blood DDE levels and urban versus rural residence and blood PCB levels in 1974. However, in 1989, none of these spatial variables were found to be significant potential risk factors for blood organochlorine levels. Evaluating these potential risk factors add statistical power to the models and are able to reduce unexplained variation.

Since spatial dependence was recognized in the residuals of the models, there is evidence that even more variation in the model can be explained by spatial information that was not explored in this study. However, since no additional spatial information was available, the best method of correcting for this dependence was recognizing it in the residuals using GLS regression methods that allow for the residual error terms to have different variances, or to be correlated (*220*).

After correcting for spatially dependent residuals, most model parameter estimates were not changed significantly. However, those covariates that bordered on statistical significance (i.e. p-values around 0.05) were sensitive to correcting for spatially dependent residuals. For example, a statistically significant urban/rural residence relationship with lipid adjusted total PCBs in 1974 was found in OLS regression, but became statistically insignificant after correcting for spatial dependence in the residuals. Furthermore, in CLUE II, age became statistically insignificant in the lipid adjusted DDE model after adjusting for spatially dependent residuals, and BMI became statistically insignificantly associated with lipid unadjusted total PCBs.

5.6. Discussion

This paper demonstrates the importance of evaluating potential spatial risk factors and taking into account residual spatial dependence in regression models attempting to explain levels of contaminants in humans. Taking into account spatial information is more commonly used in evaluating environmental contamination, but often overlooked in studies modeling the same contaminants in humans, despite the fact that biomarkers are indicators of exposure.

The results of this paper suggest that residential location may be a potential exposure determinant of organochlorine levels in human blood as biomarkers of exposure to persistent organochlorine compounds in Washington County, Maryland. A significant negative association is present between blood dieldrin levels and residential distance from the Superfund site. However, an association between residential location and the potential source of exposure in the County was not found with blood DDE levels. In fact, a weak, but positive association was found between residential distance to the Superfund site and blood levels of DDE instead of the negative association suspected. One possible reason for this contradiction may be because DDE is a widespread ubiquitous compound that can be found in the blood of over 90 percent of the United States population, whereas dieldrin is not as commonly found in the environment and in human blood (*4,36*). In addition, DDT was most likely used often and all over the County before the 1970's, especially on crops in the rural areas. Therefore, there may have been multiple non-point sources of exposure to DDE in the study population. The results in this paper suggest that this widespread use of DDT may be a much larger contributor to internal dose than any increase in body burden associated with living close to the Superfund site. Dieldrin, on the other hand, had more limited use for termite control. Therefore, the site may have been the primary source of dieldrin exposure, resulting in higher blood levels for those individuals living closer to the site.

Further research is needed to determine the validity of the association between blood dieldrin levels and the Superfund site. Not only is the statistical significance of this association marginal, but the model is based on less than half the sample size than that of DDE. Furthermore, it is troubling that the model found smoking to be negatively associated with blood dieldrin levels. There have not been any other studies in the literature suggesting such an association with smoking, and therefore, more research into this finding is warranted. Overall, the results are inconclusive as to whether or not there is a direct relationship between the distance to the CCC Superfund site and residential location and levels of organochlorines in the blood of the participants.

The results of this paper indicate a stronger relationship between residential location and blood organochlorine levels in 1974 than in 1989 since the potential spatial risk factors evaluated in this paper were found to be influential in modeling the levels of organochlorines in human blood in 1974 but not in 1989. There are a few possible reasons for this. First, DDT was banned in the United States in 1972, and the use of PCBs was severely restricted in the 1970's as well. Therefore, potential sources of exposure in this area decreased or may have disappeared altogether. Therefore, without potential sources of exposure, people living closer together may not have similar levels of organochlorines in their blood compared to people residing far from each other. Furthermore, since the sources of exposure existed mainly before the 1970's and the blood was drawn in 1989, many of the participants may have moved in those 15 years, and therefore, their residential location is not indicative of where they lived during peak exposure. Finally, the overall levels of these substances in human blood dropped by approximately 50 percent from 1974 to 1989 in this population. This may have made it more difficult to statistically detect associations with spatial risk factors.

It is interesting to note that when both whites and African Americans are considered in the 1974 DDE models, race strongly modifies the effect of distance to the site on blood DDE levels. Although distance to the site is not a statistically significant predictor in either African Americans or whites alone, there is a statistically significant difference in the effect of distance to the site on blood DDE levels. However, since this finding is only based on a population of 24 African Americans, the validity of this finding is questionable, and is the reason the results of only the white population are reported in this paper. The results of the model including both races as well as a more detailed discussion about the racial effect are found in Appendix B. Several other studies have also reported racial differences in blood organochlorine levels with higher levels in African Americans than whites (*4,41,130,167,210,221*).

For the most part, the potential risk factors found to be associated with blood organochlorines in this paper are consistent with the literature. For instance, other studies also found that blood organochlorine concentrations were positively associated with age, BMI, or current smoking status (*3,155,165,222*). Glynn *et al.* and Sala *et al.* also report that place of residence influences blood levels of organochlorines (*3,165*). The positive age association can be explained by the fact that older people have had a longer duration of exposure which is reflected in their body burden of organochlorines. In addition, since these compounds are lipophilic, people with larger BMI's are able to store greater amounts of organochlorines in their bodies. Smokers may have higher exposure to blood organochlorines due to the constant hand to mouth activity. It also makes sense that those people that rely on wells or springs for drinking water have higher levels of organochlorines in their blood, since the CCC site has been shown to have polluted the surrounding waterways, and the geology of the area is such that surface water is able to easily contaminate the groundwater (*8-10,170*).

There does exist some inconsistency in the literature regarding potential risk factors that are evaluated in this paper. For instance, this paper reports that men have higher levels of blood organochlorines (except for DDE in 1989) than women. Other studies have reported the same findings (*80,223*). On the other hand, this is inconsistent with several studies that either report women have higher levels than men, or no association with gender (*3,63,224*). A potential reason for this inconsistency could be the fact that gender does not always account for differences in occupation and BMI, as these are both potential confounders in this relationship.

The very weak associations between diet and blood organochlorine levels found in this study are not consistent with a majority of the literature, but have been reported (*3,155,165,189*). This is most likely because the diet data used in this paper were not collected for the purpose of studying organochlorine exposure and are incomplete. For example, the diet data is not categorized as locally caught, grown, or raised food. Despite the limitations in the diet information, the negative association found between vegetable and dairy consumption and blood DDE levels warrants further study. It would be expected that vegetable consumption would be associated with higher blood levels of organochlorines since they are grown in soil, and often spayed with pesticides. Likewise, since dairy products contain animal fat, it is expected that organochlorine compounds would also be present in these products.

An additional limitation resulting from the fact that the data analyzed in this study were not collected for the purposes of evaluating organochlorine exposure is that information on possible confounders or effect modifiers is not uniformly collected in the two CLUE studies. For example, the CLUE I questionnaire did not obtain height and weight measurements, and therefore BMI could not be taken into consideration for the CLUE I portion of this study. BMI was shown to be a predictor of blood organochlorine levels in the CLUE II analysis of this study as well as in the literature (*3,85,165,167,210*). Likewise, drinking water source information was found to be influential in the CLUE I models of blood organochlorines, but unavailable in the 1989 CLUE II study. Other information such as breastfeeding history, weight loss, and occupational and home exposure to pesticides in the literature have been shown to be significant predictors of blood organochlorine levels (*3,63,69,165,225*). Although this information was available for many of the participants in this study, it was collected more than five years from the time of blood draw. Since these data may not have been representative of the behavior at the time of blood draw, they were omitted from the analysis. Therefore, future studies should strive to collect information on all possible risk factors at the time of blood collection.

Although high residual error and low explained levels of variation in regression models are common when dealing with human populations due to human variability, they indicate that there is still unexplained uncertainty in these models. This paper indicates that spatial dependence in these residuals can account for some of this error using spatial information. It is therefore important to collect information on potential individual-level risk factors as well as all spatial risk factors when designing future studies. Additional potential risk factors that may have been helpful in this paper would have included: occupation, exposure to organochlorines, consumption of local and fatty fish, consumption of home-grown vegetables, recreational swimming in local surface waters, land use, and drinking water well location and/or source aquifer.

This study compares organochlorine concentrations measured in serum collected in 1974 to organochlorine concentrations measured in plasma collected in 1989. Plasma is whole blood, while serum is plasma minus the clotting agents, which may interfere with analytical techniques. Although this issue needs further exploration, it is estimated that the difference between the two concentration measures, when adjusted for lipids, is most likely small when compared to the difference that should be seen between the two time points (*60*). Hence, this issue would not be expected to generate great bias in this study.

This paper relies on two assumptions related to participant address information. First, it assumes that the participant's address represents their residential location during the time they were most exposed to organochlorines. In other words, if they had moved just prior to blood collection, the address information used in this study would only represent their current address, and not where they lived in the past. It is also possible that they may have had more exposure at their place of employment or recreation than at their residence. Furthermore, this paper assumes that the locations of the residences were geocoded accurately. However, this assumption is not valid since there exists positional inaccuracy associated with geocoding in GIS (*177,213,226*). Although this positional inaccuracy has been found to be insignificant in a study by Bonner *et al.*, the sensitivity of the models presented in this paper to this bias was evaluated in the third manuscript (*226*).

There are over 1,200 Superfund sites across the country that are contaminated with substances that adversely affect human health (*11*). Furthermore, Superfund sites are often located in urban areas surrounded by residences. This paper presents methods to better characterize residential location as a determinant in exposure by evaluating the relative contribution of spatial information in predicting human internal dose.

5.7. Acknowledgements

This study was supported by the National Institute for Environmental Health Sciences (NIEHS) training grant (ES 07141), the Johns Hopkins NIEHS Center in Urban Environmental Health (P30 ES 03819) and the National Institute for Occupational Safety and Health Education and Research Center training grant (T42-CCT310419).

MANUSCRIPT THREE

GEOCODING POSITIONAL BIAS AND ITS EFFECT IN ANALYZING THE SPATIAL DISTRIBUTION OF BLOOD DIELDRIN LEVELS

6. Manuscript Three: Geocoding Positional Bias and its Effect in Analyzing the Spatial Distribution of Blood Dieldrin Levels

6.1. Abstract

Geographic information systems (GIS) are commonly used in environmental and public health research in order to link exposure and disease to geographical locations. However, the process by which GIS links information, geocoding, is not designed to pinpoint exact address location but rather approximate address locations between street intersections. Positional inaccuracies occur in geocoding when an address is placed at an incorrect location along a street and further compounded when the distance from the street to the actual residence or building is considered. The purpose of this paper is to explore the extent and spatial variation in these sources of positional bias in geocoded residential locations from a study on organochlorine exposure and to assess the robustness of an epidemiological model to this positional bias.

Positional bias was quantified in this paper for 163 randomly selected geocoded residential addresses out of a dataset containing 1,323 geocoded addresses. Conditional simulation was then used to estimate the positional biases at all remaining geocoded addresses in the dataset 1,000 times. A new estimated residential location was created for each of the 1,000 estimated positional biases, and used in a model that found a statistically significant association between blood dieldrin levels and residential distance from a potential source after adjusting for age, gender, education, smoking status and drinking water source. Hence, the sensitivity of this epidemiological model to these geocoding positional biases was also tested.

The results of this paper indicate that the mean parameter estimates and their associated statistical significances changed minimally from the observed values accounting for the positional bias. The distribution of the 1,000 t-statistics corresponding to the residential distance from the potential source parameter in the epidemiological model indicated that this parameter was significant in over 90 percent of the trials. The overall mean statistical significance of the association between blood dieldrin levels and residential distance from the potential source after adjusting for age, gender, education, smoking status and drinking water source did not change. Therefore, the epidemiological model based on geocoded address information evaluated in this paper is robust to these geocoding positional biases.

However, the representativeness of this study population to other populations is unknown. Therefore, it is important that researchers acknowledge that positional inaccuracies are associated with geocoding, and take these biases into account if possible. Furthermore, before reporting significant results that are based on geocoded information, the sensitivity of their results to these biases should be tested.

6.2. Introduction

Evaluating spatial relationships and geographic determinants of health and exposure is a growing area of environmental and public health research. As a result, geographic information systems (GIS) have become increasingly popular in environmental health applications and to environmental health practitioners (*177-184,187*). Maps and map overlays are basic constructs of any GIS, which are linked with a geographic location. Geocoding is the GIS process by which an attribute or characteristic can be linked to a geographical location. Most often, this is accomplished using ordinary street addresses matched to digital basemaps within GIS, providing, for example, a set of longitude and latitude coordinates. With this information, GIS can serve as a relational database by providing a common format to link specific data at the location to additional geographical or environmental data collected in the same geographic area (*2*). Such spatial databases provide the necessary input to the various methods of spatial statistics available.

Coordinates obtained from GIS geocoded street addresses provide estimated locations that are subject to positional inaccuracies (*226*). GIS basemaps contain roads, which are represented as line segments, and road intersections. Geocoded coordinates are obtained by approximating numbered addresses in proportion to the street segments between intersections. For example, the geocoded coordinate for 125 Main Street would be the location representing one quarter of the way between the 100 and 200 blocks of Main Street, assuming the basemap contains these block intersections. The distance along the road between the true location and the geocoded coordinate is one potential source of positional bias. Bonner *et al.* evaluated this positional bias and found that 79 percent of the 200 addresses they sampled were within 100 meters of the geocoded position (*226*). The distance residences are from the street is a second potential source of positional bias with some residences being close to the street and others being set back some distance.

Note, in GIS software, such as ArcGIS (Environmental Systems Research Institute, Inc.), users are allowed to specify what is called an offset in the geocoding process (*201*). This offset quantity is designed to relocate the geocoded coordinates from the middle of the street, which is commonly the default process, to one side of the street by a distance specified in the offset. Since it will be shown that the distance residences are from the street as well as the other sources of geocoding positional bias vary spatially, it is not clear if this geocoding offset mechanism can be useful in minimizing the positional bias considered in this paper.

The extent and possible spatial variation in the positional accuracy of geocoded coordinates is of important interest. In environmental and public health research applications, geocoded addresses often represent a proxy for exposure. Statistical models for exposure assessment in such studies that use geographic information are therefore subject to potential misclassification or bias due to these inaccuracies in location information (*187*). Using universal kriging in the second manuscript entitled, "The Use of Spatial Information in Determining Potential Risk Factors of Blood Organochlorine Levels in a Population Living Near a Potential Source," blood levels of the insecticide dieldrin were found to be significantly associated with the distance subjects lived from a former pesticide facility. The distance to this source was calculated based on geocoded location information, and therefore, depends on the positional accuracy of the geocoded coordinates. Hence, the question is raised as to how robust these results are with respect to geocoding positional bias. Even in situations when a location-derived potential risk factor (e.g. distance to source) is not considered, spatial regression models such as kriging rely heavily on location information for characterizing residual spatial variation, and would also be subject to this bias.

There are two objectives to this paper. The first objective is to explore the extent and spatial variation in the positional bias of the geocoded residential locations from a study on organochlorine exposure. A global positioning system (GPS) and a laser distance meter device are used to measure both potential sources of positional bias. Second, the positional biases obtained from this sample are used to assess model robustness in the cited dieldrin study that included distance to a potential source as a significant effect.

6.3. Blood and Geocoding Positional Bias Datasets

In 1962, the Johns Hopkins Training Center for Public Health Research was founded in western Maryland. Among its many accomplishments, the Center sponsored the collection of two large blood specimen banks in 1974 and 1989 (*164*). Blood samples were assayed for 1,1,1-trichloro-2,2-bis(*p*-chlorophenyl)ethane (DDT), 1,1-dichloro-2,2-*bis*(*p*-chlorophenyl)ethylene (DDE), and 28 polychlorinated biphenyl (PCB) congeners. A subset of these samples were also analyzed for additional organochlorines such as dieldrin (*60,131,148*). Details concerning the blood collection, storage, and analytical methods are found in the background section and published elsewhere (*60,131,148*). Several analyses of these data exist (*60,131,148*). Of interest to this paper is a study that found residential proximity to a former pesticide facility to be a significant potential risk factor for blood dieldrin levels in 1974 when adjusting for age, gender, education, smoking status, and drinking water source. This study is found in the second manuscript entitled, "The Use of Spatial Information in Determining Potential Risk Factors of Blood Organochlorine Levels in a Population Living near a Potential Source."

The pesticide facility, in operation at the time of the first blood sample collection (1974), has since been placed on the National Priority List as a Superfund site (1997)

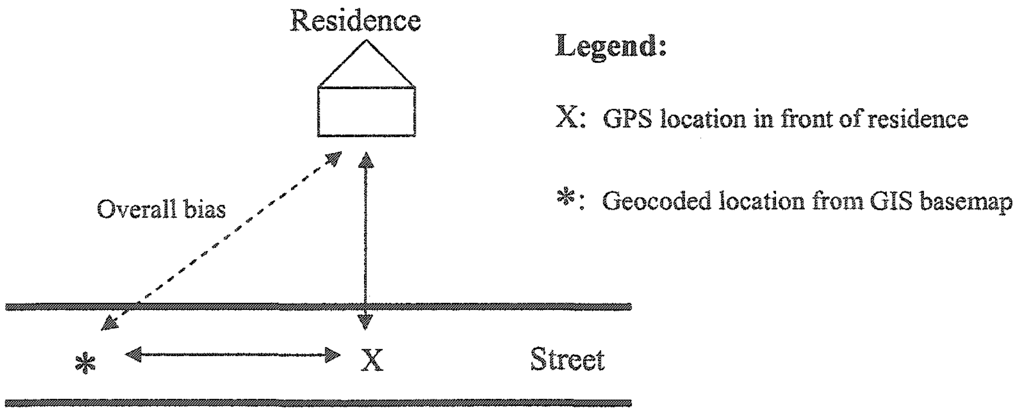
(*11,194*). The 19-acre site was built in the early 1930's by a large chemical company for the production of fertilizers and pesticides, including DDT and dieldrin, which it did until 1968. DDT and its derivative, DDE, as well as dieldrin are classified by the United States Environmental Protection Agency (EPA) to be Group B2, probable human carcinogens (*6,227*). Acute doses of these compounds have also been found to severely affect the nervous system (*6,36,227*). In 1972, the use and production of DDT was banned in the United States. The EPA banned dieldrin for all uses except to control termites in 1974, and banned it completely in 1987.

The database of blood organochlorine levels included geocoded addresses for 1,323 participants, as well as individual characteristics such as age, gender, education level, smoking status, and drinking water source. The geocoded address coordinates and the coordinates corresponding to the center of the Superfund site (obtained using a Garmin model V GPS) were used to calculate distance to the site. The geocoded coordinates were obtained in ArcGIS using StreetMap 2000 (Environmental Systems Research Institute, Inc.) and United States Census Bureau 2000 Topologically Integrated Geographic Encoding and Referencing (TIGER) files as basemaps. Details regarding the geocoding process are found in the research methods section.

A stratified random sample of 200 of these geocoded addresses was selected for measuring positional bias, 100 located in an urban setting and 100 from a more rural environment. Urban residence is defined in this paper as those residents whose zip code corresponded to one of the two zip codes associated with the largest municipality in the County. All other addresses were designated as rural.

During a visit to each address, GPS coordinates were obtained from the center of the street in front of the house, denoted as position "X" in Figure 6.1. The Euclidian distance between the GPS coordinates (X) and the corresponding geocoded coordinates (*) was calculated to determine one source of positional bias. The distance from this position (X) to the actual residence was then measured using a laser distance meter device (CST/berger model LS-1), capturing another source of geocoding positional bias. If the distance from the house to the street was greater than 100 meters (328.1 feet), the laser meter was unable to capture it, and 328.1 feet was recorded. The actual or overall positional bias, the distance from the geocoded coordinates to the physical residence (represented as a dashed line) was then calculated assuming measurements were made in a perpendicular fashion as shown in Figure 6.1.

Figure 6.1: Diagram of positional biases associated with geocoding a residence in GIS


Residence
Legend:
X: GPS location in front of residence
*: Geocoded location from GIS basemap
Overall bias
*
X
Street

Out of the 200 residences that were randomly chosen to visit, a total of 163 residences were successfully physically located and positional data obtained. The remaining residences were unable to be located and/or measured because either the houses no longer existed or were inaccessible. Overall, 85 urban and 78 rural residences were available for analysis.

6.4. Variation in Geocoding Positional Bias

A summary of the two sources of geocoding positional bias as well as the overall bias calculated from the two sources (Figure 6.1) is displayed in Table 6.1. As could be expected, the distance from the house to the street, on average, was higher for rural residences than urban (88 versus 60 feet). In addition, there was only one home that was beyond 328.1 feet from the street in the urban area and seven in the rural areas. Likewise, the distance from the geocoded coordinates to the actual coordinates was greater in the rural zip codes than in the urban zip codes. Hence, the overall median positional bias was also higher for rural residences than urban residences (318 versus 285 feet). This overall geocoding positional bias combining both urban and rural data is used for further analysis.

Table 6.1: Summary of Measured Geocoding Positional Bias Results in Washington County, Maryland

	Median Distance (feet)	Mean Distance (feet)	Minimum Distance (feet)	1st Quartile Distance (feet)	3rd Quartile Distance (feet)	Maximum Distance (feet)	Number beyond 328 feet
Residence to Street							
Total (N=168)	59	74	16	47	74	328	8
Urban (N=88)	56	60	18	42	67	328	1
Rural (N=80)	66	88	16	50	81	328	7
Geocoded to GPS Coordinates							
Total (N=168)	295	1,385	19	125	612	26,400	NA
Urban (N=90)	290	1,633	31	135	629	22,180	NA
Rural (N=78)	302	1,100	19	116	597	26,400	NA
Residence to Geocoded Coordinates							
Total (N=163)	302	1,409	44	141	608	26,400	NA
Urban (N=85)	285	1,673	44	147	601	22,180	NA
Rural (N=78)	318	1,121	55	130	610	26,400	NA

NA: Not applicable

The positional bias data were skewed and assumed to be distributed lognormally, as shown in Figure 6.2. Therefore, biases were log transformed for all analyses, but back transformed for reported results. An exponential spatially dependent correlation structure was fit to the geocoding positional bias data using the maximum likelihood method, and is shown in Figure 6.3 (*23*). From this variogram, it appears that positional biases are no longer spatially correlated beyond approximately four miles.

Figure 6.2: Histogram of geocoding positional bias data measured at 163 residential locations in Washington County, Maryland

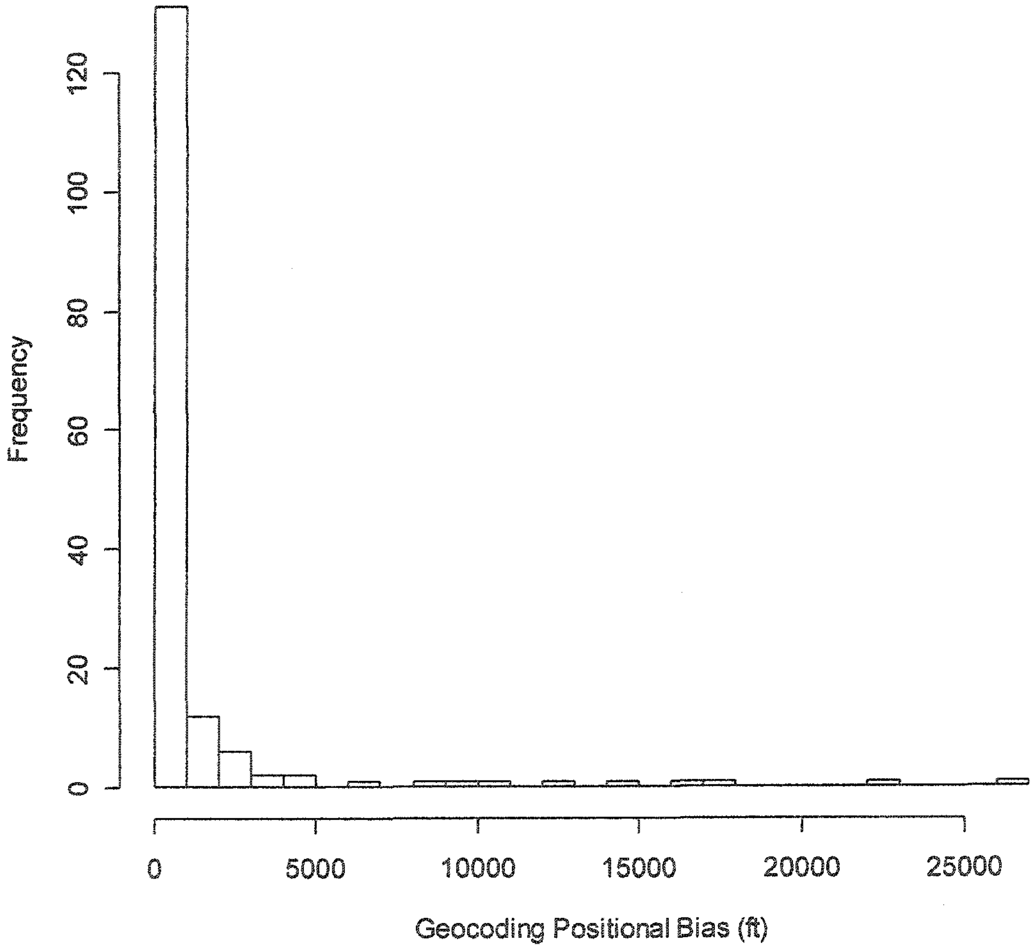
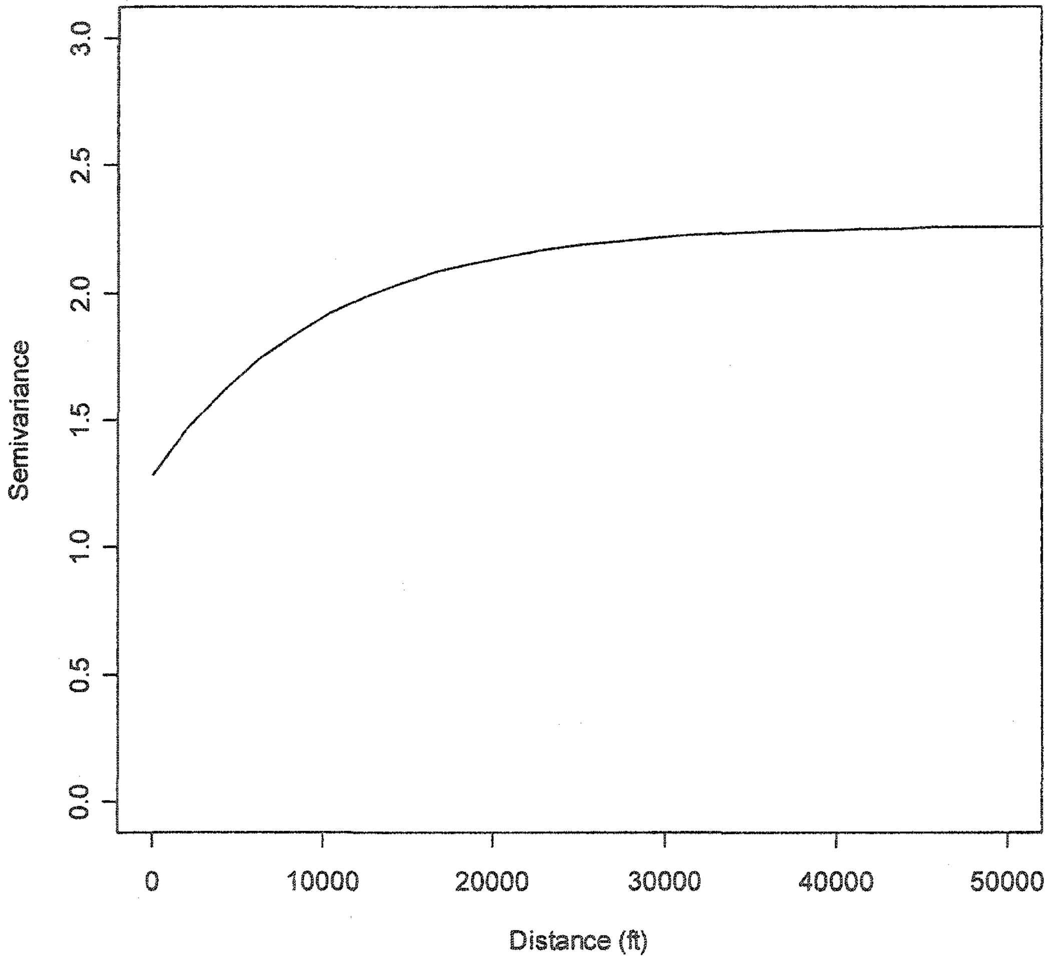


Figure 6.3: Exponential spatial dependence correlation model fit to geocoding positional bias data measured at 163 residential locations in Washington County, Maryland



6.5. Effect of Geocoding Positional Bias in Kriging Models

The regression model used to analyze the blood dieldrin data in the second manuscript is of the form:

$$Y(\mathbf{s}) = \beta_0 + \beta_1 X_1(\mathbf{s}) + \ldots + \beta_n X_n(\mathbf{s}) + \varepsilon(\mathbf{s}), \tag{6.1}$$

where '$\mathbf{s}$' denotes the geocoded residential coordinates, $Y(\mathbf{s})$ represents blood dieldrin levels for participants residing at location $\mathbf{s}$, $X_1(\mathbf{s})\ldots X_n(\mathbf{s})$'s risk factors indexed by location $\mathbf{s}$, $\beta_1\ldots\beta_n$'s, their associated effects, and β_0 the baseline intercept. The residual error term, $\varepsilon(\mathbf{s})$ was assumed normally distributed with a zero mean and constant variance. To further account for possible residual spatial variation, residuals were allowed to be spatially dependent by parameterizing their correlation as a decreasing function of the distance between their locations. In the geostatistical literature, Model 6.1, with these specifications, is known as a universal kriging model commonly used for spatial prediction at unobserved or unmeasured locations (*23*). The goal in the second manuscript was more inferential with a primary focus in selecting and quantifying potential risk factors that best characterize blood organochlorine levels.

The final model arrived at in the second manuscript for blood dieldrin included significant effects due to residential distance to the Superfund site, current smoking status (yes/no) and drinking water source (municipal/well/spring) adjusting for these covariates as well as age and gender. A summary of the covariate effects and their corresponding p-values are reproduced in Table 6.2. Inferences for the model parameters were obtained using the following two stage approach (*23*). Variogram functions were estimated for the ordinary least squares (OLS) residuals from Model 6.1, and used to specify the residual correlation structure. The residual correlation structure was then used to re-estimate model effects using generalized least squares (GLS) methods (*190*). To reduce bias from using an estimated correlation structure based on the OLS model, another variogram was produced based on the GLS residuals, and a new correlation structure was estimated. This structure was then used to further update the GLS estimated effects, completing the two stage approach. Standard errors and t-tests for statistical significance were then generated.

Of primary interest in this paper is the effect of geocoding positional bias on the estimation and significance testing of the distance to the Superfund site risk factor. Geocoding positional bias directly effects distance calculations, which in Model 6.1 involve the distance to the Superfund site variable as well as the residual error component that is assumed spatially dependent. Sensitivity of either of these model components to this bias could impact the results. Note the remaining potential risk factors in the model are individual characteristics and are most likely robust with respect to geocoding positional bias. Any changes in their effects would only be due to a different dependent error term.

An approach based on conditional simulation is used to assess the robustness of the dieldrin universal kriging model results to geocoding positional bias (*23,228*). Let $S = \{s_1,\ldots,s_{443}\}$ represent the set of geocoded coordinates for all the blood dieldrin data used to fit Model 6.1 in the second manuscript. Let $S^*=\{S_1^*, S_2^*\}$ represent the complete set of "true" coordinates with $S_1^*=\{s_1^*,\ldots,s_{163}^*\}$ denoting the subset corresponding to the sample geocoded addresses that were visited and S_2^* the remaining 280 true coordinates. Following this notation, let $R=\{R_1, R_2\}$ represent the complete set of geocoded positional biases as those in S_1^* are taken here to represent those obtained after relocating the corresponding geocoded coordinates by their measured positional bias in R in a random direction. R_1 was observed, and used to generate S_1*. Conditional simulation is used to generate positional bias information for the unsampled addresses referred to as R_2 thus providing the necessary information for generating true coordinates in the remaining subset, S_2*.

The positional bias, R_1, explored previously, was assumed to be lognormally distributed. Conditional simulation for positional bias in R_2 is simulated from the conditional distribution of R_2 given the already measured positional bias in R_1. Assuming the complete (R_1, R_2) distribution of geocoding positional bias follows that of the subset R_1 which was lognormal, the conditional distribution of R_2, given R_1, can be specified and used to produce simulated positional biases. There are two main advantages of using the conditional simulation approach. First, the resulting simulated surface passes through, or is conditioned on, the observed set of positional biases in R_1. This is ideal since these were actually measured in the field. In other words, the positional biases measured in R_1 remain fixed. Second, the spatial structure of the positional bias estimated with the observed set, R_1, (the variogram in Figure 6.3) is preserved with the conditionally simulated R_2 data. The following algorithm was used with conditional simulation to assess the robustness of model results to geocoding positional bias.

1. Estimate and model spatial dependence for the sampled geocoding positional bias, R_1 (Figure 6.3).
2. Based on results from Step 1, conditionally simulate geocoding positional bias for the remaining geocoded addresses in the blood organochlorine database, R_2.
3. Generate new residential location, S*, based on the measured positional bias, R_1, and the simulated positional bias, R_2, and from this, a new distance to the Superfund site variable.
4. Re-estimate Model 6.1 using the two stage GLS approach outlined above using the new residential locations, S*, and distance to the Superfund site variables.
5. Repeat Steps 2 through 4 1,000 times.
6. Summarize the distribution of re-estimated distance to the Superfund site effects and their t-statistics.

The exponential variogram model was fixed throughout the simulations, and estimated with maximum likelihood (Figure 6.3). Various checks at different stages in the process supported this fixed variogram model. All statistical analysis was done using the open source R statistical computing environment with the contributed package, geoR, for spatial statistical operations (*31,192*).

The distributions of the 1,000 beta coefficients and t-statistics for the distance to the site variable taking into account geocoding positional biases in the model of blood dieldrin levels after adjusting for age, gender, education, smoking status, and drinking water source are found in Figures 6.4 and 6.5, respectively. The observed beta coefficient and p-value in the same model not taking geocoding positional bias into account from the second manuscript are marked on the distributions with a line. These lines appear to be located in the center of the distributions. Therefore, the regression model from the second manuscript is not sensitive to the geocoding positional bias quantified in this paper. The distribution of t-statistics resulting from the conditional simulation algorithm shows that the effect of the distance to the site variable on blood dieldrin levels was statistically significant in over 90 percent of the trials.

Table 6.2 shows the mean beta coefficients and p-values obtained from the mean t-statistics for all variables in the model and compared to the observed values from the second manuscript that did not take geocoding positional bias into account. Overall, the beta coefficients in the model changed minimally, not affecting the t-statistics, and therefore leaving the p-values unchanged. Hence, taking geocoding positional bias into account did not change the overall results reported in the second manuscript. More details regarding the model results are found in the second manuscript.

Figure 6.4: Distribution of 1,000 beta coefficients of a residential distance to potential exposure source variable taking geocoding positional bias into account in an epidemiological model of blood dieldrin levels adjusted for age, gender, education, smoking status and drinking water source and compared to the beta coefficient value of the same variable in the same model not taking geocoding positional biases into account (represented as a line)

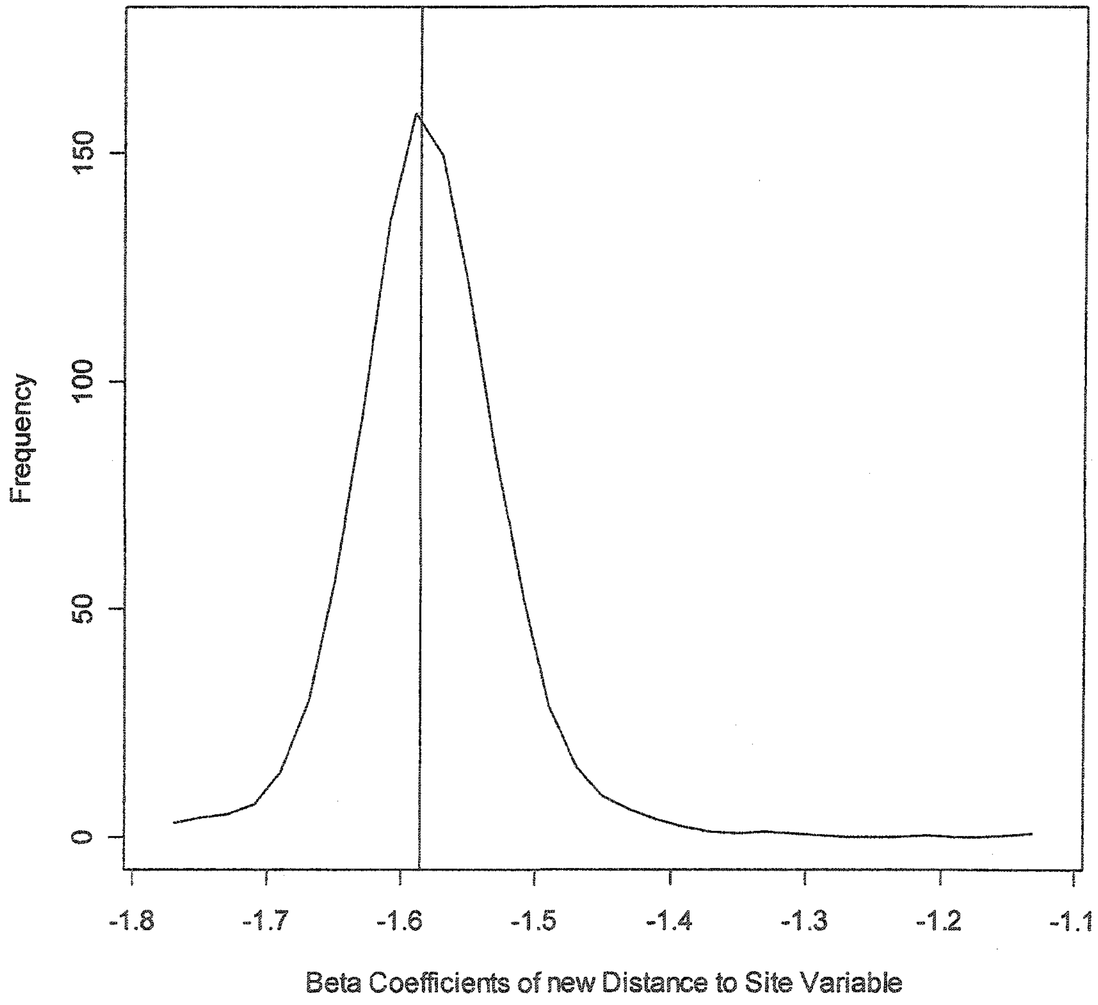


Figure 6.5: Distribution of 1,000 t-statistics of a residential distance to potential exposure source variable taking geocoding positional bias into account in an epidemiological model of blood dieldrin levels adjusted for age, gender, education, smoking status and drinking water source and compared to the beta coefficient value of the same variable in the same model not taking geocoding positional biases into account (represented as a line)

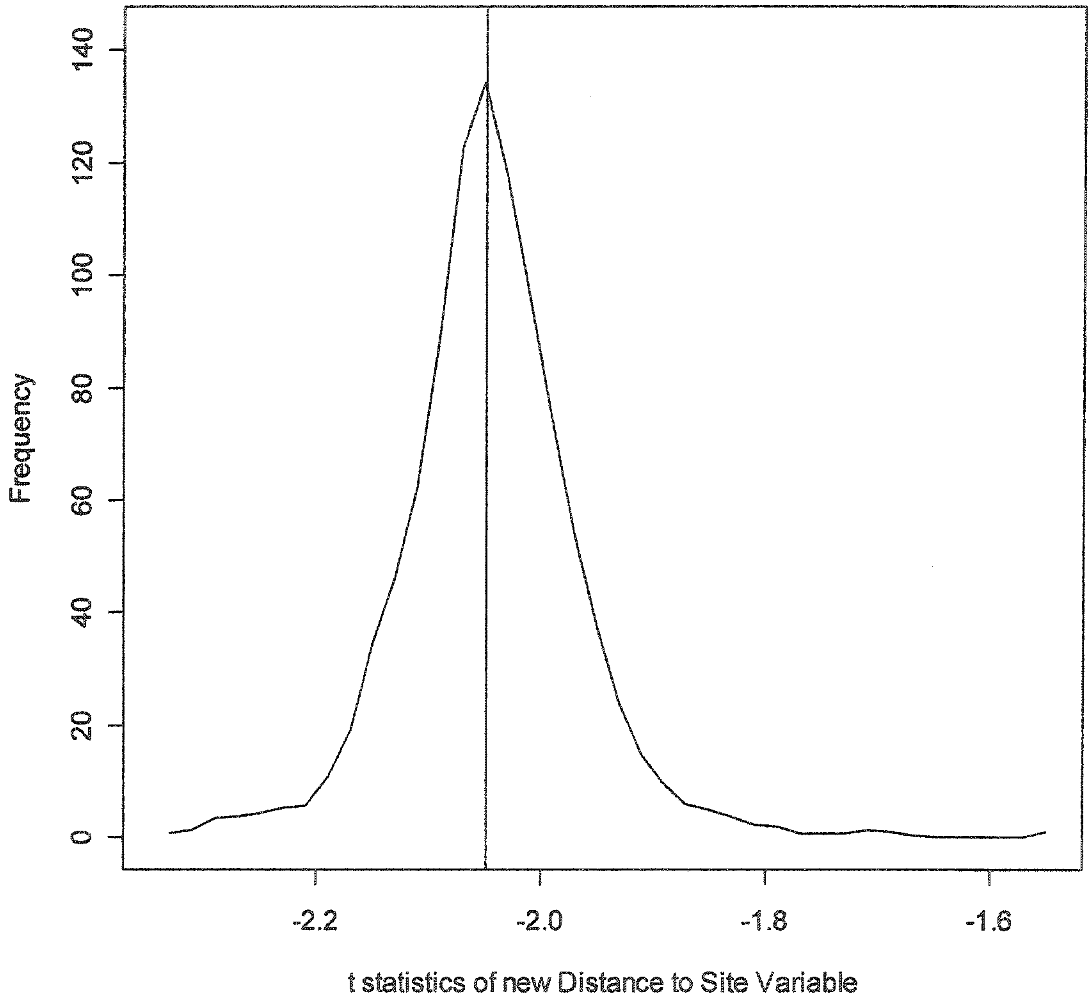


Table 6.2: Results of Multivariate Linear Regression Determining the Effects of Covariates on Lipid Adjusted Blood Dieldrin Levels Both With and Without Taking Geocoding Positional Bias into Account: Parameter coefficients (p-values)

Covariate	Observed Parameter Estimates without Accounting for Geocoding Error	Expected Parameter Estimates Accounting for Geocoding Error
Distance from CCC site to Residence (miles)	-1.5867 (0.042)	-1.5794 (0.042)
Age (years)	-0.1894 (0.5)	-0.1890 (0.5)
Gender (male vs. female)	14.5878 (0.09)	14.6009 (0.09)
Education (years)	-1.7726 (0.2)	-1.7685 (0.2)
Smoking Status (non-smoker vs. smoker)	21.7963 (0.004)	21.7673 (0.004)
Drinking water Source (spring vs. municipal)	221.1706 (<0.001)	221.1256 (<0.001)
Drinking water Source (well vs. municipal)	9.3924 (0.4)	9.3829(0.4)

6.6. Discussion

Geocoding positional biases occur for many reasons. First, and foremost, the accuracy of geocoding is strongly dependent on the accuracy of the basemap being used. The basemap should be up to date, and contain all streets. However, even the most current basemaps such as those provided with GIS systems and Census TIGER files can't capture all of the side streets and rural routes (*2,177*). It is therefore recommended that several basemaps be used to minimize the number of missing streets (*177,213*). Furthermore, in addition to the bias associated with matching addresses based on a proportional distance between intersections, the same sort of bias could occur if the houses on the block do not take up equal proportions of the line segment. For instance, there could be several small businesses next to one large building that takes up over half the block. The GIS system is unable to recognize this unequal proportion, and may place the address in an inaccurate location (*177*).

In this paper, the average distance from the geocoded to GPS coordinates was found to be very large. This is inconsistent with a study by Bonner *et al.* who reported that the positional inaccuracy was relatively insignificant in their study conducted in Western New York State (*226*). The reasons for this are unclear. It is possible that the basemaps were simply not accurate for Washington County's relatively rural location. In addition, several of the residences were on long rural streets with few intersections and few houses. This may enlarge the bias associated with placing addresses at the incorrect location based on the street address between intersections, or blocks. Another potential source of the large distances could be that the researchers located and measured the incorrect house. Many of the houses and streets were not marked well, and therefore, it is possible, although unlikely, that a mistake could have been made. Nonetheless, this issue warrants further study.

Measuring the distance from the street to the actual house is novel in the literature. Although it is an obvious geocoding positional bias, it is more difficult to measure, and is probably the reason it has not been evaluated in previous studies. Although GIS software, such as ArcGIS, allows the user to estimate this distance (offset) from the center of the street, the user is confined to defining only one constant distance on one side of the street, or manually entering in an offset for each address. As evidenced by this paper, these distances are not constant, and rather vary spatially, thereby limiting the effectiveness of this GIS tool. However, with new laser technology that measures relatively large distances, researchers can measure the distance from the street to the building without having to enter the property. Future researchers evaluating geocoded addresses should take this into consideration.

It was expected that there would be a difference in geocoding positional bias between urban and rural residences (*184,229*). However, the fact that the difference was minimal was most likely due to the way this paper defined urban residence. Although it made sense to define urban by the zip codes that correspond to the largest municipality in the County, when the locations were visited, there were several residences that appeared to be rural despite the fact that they had an urban zip code and vice versa. Perhaps a better definition of urban versus rural would have been based on population density as was done in a study by McElroy *et al* (*177*).

The positional bias measured in this paper is only one source of potential bias to studies that involve geocoded information. Another limitation is the fact that not all addresses are able to be geocoded. Post office boxes and rural routes, for instance, are not recognized by many basemaps (*178*). The reason why addresses were not geocoded is important and should raise concerns. If all addresses that weren't able to be geocoded were located in the same region, or were similar demographically or biologically, omitting them from analysis could bias results (*187*). In the literature, an attempt to minimize this bias has been to either trace down the correct address, or geocode the address to a larger area, such as a census tract or zip code, as opposed to exact address location (*178,180,186,187,230*). However, tracing addresses can be very difficult and resource consuming, and geocoding to a larger area introduces ecological fallacy (*178,180,186,187,230*).

The results of this paper show that taking geocoding positional bias into account may not affect the results of studies that rely on geocoded information in Washington County, Maryland. Specifically, the results of this paper strengthen the results reported in the second manuscript that suggest blood dieldrin levels are influenced by residential proximity to a potential source of exposure. Since GIS and geocoding are becoming more popular in environmental and public health research as technology advances, it is important that researchers acknowledge that positional biases are inherent in geocoding, and take these biases into account if possible. Furthermore, the sensitivity of their results to these biases should also be tested, before reporting significant findings that are based on geocoded information.

6.7. Acknowledgements

This study was supported by the National Institute for Environmental Health Sciences (NIEHS) training grant (ES 07141),the Johns Hopkins NIEHS Center in Urban Environmental Health (P30 ES 03819) and the National Institute for Occupational Safety and Health Education and Research Center training grant (T42-CCT310419).

CONCLUSIONS

7. Conclusions

There have been many studies characterizing the geographic distribution as well as the transport and transformation of organochlorine compounds in the environment (*1,195*). In addition, there have been numerous studies evaluating the potential adverse health effects caused by organochlorines (*36*). Fewer studies have suggested how these chemicals get from the environment into the body, and most of these use biomarkers of organochlorine exposure to identify high risk factors for such routes of transmission. Although residential location near a point source has been suggested to influence exposure, spatial or geographic attributes of biomarkers of such exposure have not been well studied (*3,4*). Therefore, this research evaluates residential location as a possible exposure determinant for elevated biomarker concentrations. Spatial distribution characteristics of human blood organochlorine levels in a community with a potential source, Washington County, were evaluated in this research. The results presented in this thesis add to the growing literature evaluating the importance of geographic location as an exposure determinant. This study is one of the first to look at the spatial characteristics of biomarkers of organochlorine exposure.

7.1. Overall Findings

The results of this research highlight the importance of taking spatial information into account when dealing with both human exposure and human internal dose measures of pollutants, such as organochlorines. The first manuscript entitled, "GIS and Geostatistics: Tools for Characterizing Environmental Contamination" demonstrates the utility of Geographic Information Systems (GIS) and the R Statistical Computing

Environment as it characterizes environmental contamination around a Superfund site that is currently becoming surrounded by residences. The spatial statistical methods of variogram analysis and kriging are described and used to evaluate levels of environmental contamination. The results of this study show that estimated levels of 1,1-dichloro-2,2-*bis*(chlorophenyl)ethylene (DDE) in soil surrounding the Superfund site are well above background levels, thereby supporting the hypothesis that the Superfund site is a potential source of organochlorine exposure to the surrounding County.

Since the Superfund site is a potential source of exposure to Washington County residents, the second manuscript entitled, "The Use of Spatial Information in Determining Potential Risk Factors of Blood Organochlorine Levels in a Population Living Near a Potential Source" investigates whether or not there is a relationship between place of residence and levels of DDE, total polychlorinated biphenyls (PCBs), and dieldrin in the blood of these residents. In this paper, spatial information is incorporated in regression models that take into account residual spatial variation. Results indicate that spatial explanatory variables such as residential distance to the Superfund site and urban/rural residence may be potential risk factors for elevated levels of blood organochlorines. The study found that there was a statistically significant association between lipid adjusted blood dieldrin levels in 1974 and residential distance from the Superfund site, suggesting that those living closer to the site have higher levels of dieldrin in their blood than those living further away. However, significant associations between residential distance from the site and blood levels of other organochlorines used at the site were not found. In 1989, spatial relationships did not exist for any compounds except for an insignificant urban/rural residence association with lipid adjusted total PCBs in blood.

The study also found that assuming independent residuals in regression models when dealing with exposure models is not always a valid assumption. Generalized least squares regression techniques should be used in place of ordinary least squares as a preferred approach.

The results of the second manuscript support the hypothesis that spatial information can further predict individual blood organochlorine biomarker levels. In addition, the hypothesis that the potential risk factors for, and spatial distribution of human blood organochlorine biomarker levels in Washington County will differ in 1974 compared to 1989 is supported by this study. However, the hypothesis that there will be a spatial relationship between the blood levels of those organochlorines produced/used by the CCC and the Superfund site location is not well supported by the results of the second manuscript since it was only found with one compound at one point in time.

Results of the second manuscript are strongly based on geocoded residential information. Therefore, the sensitivity of these models to inaccuracies in the geocoding process was investigated in the third manuscript entitled, "Geocoding Positional Bias and its Effect in Analyzing the Spatial Distribution of Blood Dieldrin Levels." Two types of geocoding positional bias were measured at a random subset of the geocoded addresses. These biases were then used to predict the positional bias associated with all the other addresses that were geocoded, but not part of the subset that was physically measured. This was done using conditional simulation so that the sampled set of positional biases were fixed and the spatial dependence structure estimated from the sampled biases was preserved. These simulated positional biases were then incorporated into the 1974 blood organochlorine models from the second manuscript to assess model robustness.

The results of this study found that although geocoding positional bias was relatively large in some cases, it did not modify the regression models. Therefore, the hypothesis that geocoding positional bias will not significantly influence the regression models used to describe the blood organochlorine levels was supported.

7.2. Further Research

There are numerous avenues of potential future research to follow up these study findings. Many of these avenues stem from limitations in this thesis that can be attributed to the fact that much of the data used in this study were not collected for the purposes of this thesis. Detailed discussions regarding these limitations are found within each of the three manuscripts.

Using this research as preliminary work, a state-of-the art study could be designed to evaluate the geographic distribution of other types of environmental pollutants that pose a health threat to human populations. Focusing on geographic determinants of environmental pollutants, this future study can provide a better understanding of the pathway from source to health outcome. This information can then be used to recommend optimal control/intervention strategies. The results of this thesis suggest that this goal can best be accomplished by spatially linking environmental contamination to human biomarker data. All potential sources of the contaminant should be fully characterized spatially and temporally around the study population. Environmental samples in all relevant media (air, water, soil, and biota) should be collected throughout the region of interest. Extensive environmental sampling would be especially important surrounding potential sources of the contamination and residential, recreational, or otherwise highly populated areas. A sampling strategy designed to evaluate spatial variability would need to be developed. The levels of contamination at any unsampled locations could be predicted throughout the region using the kriging methods presented in this preliminary research. From these data, extensive environmental exposure maps could be created. In combination with time activity data, these maps could predict human exposure at different areas throughout the region which could further be used to model human internal dose measures of these pollutants.

In order to understand human exposure to environmental contaminates, potential risk factors of exposure as well as internal dose measures should be measured in a large sample of the population. In addition to the potential risk factors evaluated in this preliminary research, other potential risk factors such as occupational and personal exposure to the pollutant of interest, breastfeeding history, consumption of locally caught fish and meat and locally grown fruits and vegetables should be collected at the time the biomarker specimen is taken. In addition, spatially dependent potential risk factors such as current and historical land use, primary wind and groundwater direction, and drinking water aquifer source could also be helpful in conjunction with the potential spatial risk factors studied in this preliminary research. Taking into account this additional spatial information is needed to evaluate more complex spatial relationships than are evaluated in this preliminary work.

Furthermore, studying the spatial attributes of the health outcomes associated with exposure to these pollutants would help complete the future study. Historical as well as prospective medical records of study participants should be collected in order to evaluate the spatial relationships or patterns in these health outcome data. With all of this information, the future study would be able to more fully evaluate the potential spatial links between pollutant sources, environmental contamination and exposure, human internal doses of these compounds, and their effects on human health.

7.3. Public Health Significance

Currently, there are over 1,200 Superfund sites across the country that are contaminated with substances that adversely affect human health (*11*). Furthermore, Superfund sites are often located in urban areas surrounded by residences. Therefore, the health of people living near Superfund sites is dependant on the ability to accurately characterize and evaluate environmental contamination on and around these sites. Research is lacking in assessing the impact of residing near a potential pollution source such as Superfund sites. The results of this study help to better characterize residential location as a determinant in organochlorine exposure by determining the relative contribution of spatial information in predicting internal dose around one site.

In general, this study provides insight into additional means of predicting and evaluating human exposure. The use of spatial information helps to better characterize the pathway from source of exposure to human health outcomes. By characterizing this pathway, both spatial and nonspatial risk factors for exposure can be identified. This knowledge is needed to perform risk assessments and to design intervention strategies. Adding spatial risk factors for exposure into risk assessment techniques will help reduce the inherent uncertainty associated with assessing risk in a human population. These methods could further be used in policy decisions to implement environmental controls and public health methods of intervention that may limit human disease caused by pollutant exposure. They may also be helpful in decisions regarding such things as exposure regulations, Superfund site remediation, dietary recommendations, pesticide application, and even drinking water standards.

REFERENCES

8. References

1. Inter-Organization Programme for the Sound Management of Chemicals. Persistent Organic Polutants Assessment Report. PCS/95.38 World Health Organization (1995).
2. Vine MF, Degnan D, Hanchette C. Geographic information systems: their use in environmental epidemiologic research. Environ Health Perspect 105: 598-605 (1997).
3. Sala M, Sunyer J, Otero R, Santiago-Silva M, Camps C, Grimalt J. Organochlorine in the serum of inhabitants living near an electrochemical factory. Occup Environ Med 56: 152-158 (1999).
4. Department of Health and Human Services CfDCaP. Second National Report on Human Exposure to Environmental Chemicals. 02-0716 Atlanta: National Center for Environmental Health, Division of Laboratory Sciences (2003).
5. Holoubek I, Kocan A, Holoubkova I, Hilscherova K, Kohoutek J, Falandysz J, Roots O. Persistent, Bioaccumulative and Toxic Chemicals in Central and Eastern European Countries - State-of-the-art Report. Brno, Czech Republic: RECETOX - TOCOEN & Associates (2000).
6. Agency for Toxic Substances and Disease Registry. ToxFAQs for DDT, DDE, and DDD. Available: http://www.atsdr.cdc.gov/tfacts35.html (2003).
7. Maryland Department of the Environment. Expanded Site Inspection of the Central Chemical - Hagerstown Site - MD 302. 1 (1996).

8. Maryland Department of the Environment. Sampling Proposal for Expanded Site Inspection of the Central Chemical - Hagerstown Site MD-302 (1993).

9. Maryland Department of the Environment. Expanded Site Inspection of the Central Chemical - Hagerstown Site -MD 302. 1 (1994).

10. Maryland Department of the Environment. Expanded Site Inspection Sampling Plan of the Central Chemical- Hagerstown Site MD 302 (1995).

11. United States Environmental Protection Agency. Final National Priorities List Sites. Available: http://www.epa.gov/superfund/sites/query/queryhtm/nplfin1.htm (2003).

12. Lang L. GIS for Health Organizations. Redlands: Environmental Systems Research Institute, Inc. (2000).

13. Fitzgerald EF, Brix KA, Deres DA, Hwang SA, Bush B, Lambert G, Tarbell A. Polychlorinated biphenyl (PCB) and dichlorodiphenyl dichloroethylene (DDE) exposure among Native American men from contaminated Great Lakes fish and wildlife. Toxicol Ind Health 12: 361-368 (1996).

14. Shone SM, Ferrao PN, Lesser CR, Norris DE, Glass GE. Analysis of mosquito vector species abundances in Maryland using geographic information systems. Ann N Y Acad Sci 951: 364-368 (2001).

15. McKee Jr KT, Shields TM, Jenkins PR, Zenilman JM, Glass GE. Application of a geographic information system to the tracking and control of an outbreak of shigellosis. Clin Infect Dis 31: 728-733 (2000).

16. Latkin C, Glass GE, Duncan T. Using geographic information systems to assess spatial patterns of drug use, selection bias and attrition among a sample of injection drug users. Drug Alcohol Depend 50: 167-175 (1998).

17. Glass GE, Amerasinghe FP, Morgan JM, III, Scott TW. Predicting Ixodes scapularis abundance on white-tailed deer using geographic information systems. Am J Trop Med Hyg 51: 538-544 (1994).

18. Glass GE, Schwartz BS, Morgan JM, III, Johnson DT, Noy PM, Israel E. Environmental risk factors for Lyme disease identified with geographic information systems. Am J Public Health 85: 944-948 (1995).

19. Becker KM, Glass GE, Brathwaite W, Zenilman JM. Geographic epidemiology of gonorrhea in Baltimore, Maryland, using a geographic information system. Am J Epidemiol 147: 709-716 (1998).

20. Aronoff S. Geographic Information Systems: A Management Perspective. Ottawa: WDL Publications (1989).

21. Cressie, N. Geostatistics. The American Statistician 43: 197-202 (1989).

22. Isaaks EH, Srivastava RM. An Introduction to Applied Geostatistics. New York: Oxford University Press (1989).

23. Cressie N. Statistics for Spatial Data. New York: John Wiley & Sons, Inc. (1991).

24. Juang KW, Lee DY, Ellsworth TR. Using rank-order geostatistics for spatial interpolation of highly skewed data in a heavy-metal contaminated site. J Environ Qual 30: 894-903 (2001).

25. Saito H, Goovaerts P. Selective remediation of contaminated sites using a two-level multiphase strategy and geostatistics. Environ Sci Technol 37: 1912-1918 (2003).

26. Thayer WC, Griffith DA, Goodrum PE, Diamond GL, Hassett JM. Application of geostatistics to risk assessment. Risk Anal 23: 945-960 (2003).

27. Johnston K, Ver Hoef JM, Krivoruchko K, Lucas N. Using ArcGIS Geostatistical Analyst. Redlands: Environmental Systems Research Institute (2001).

28. Ihaka R, Gentleman R. R: A Language for Data Analysis and Graphics. Journal of Computational and Graphical Statistics 5: 299-314 (1996).

29. Deutsch CV, Journel AG. GSLIB Geostatistical Software Library and User's Guide. New York: Oxford University Press, Inc. (1992).

30. Mathsoft. S+SPATIALSTATS User's Manual. Seatle: Mathsoft, Inc. (1996).

31. R Development Core Team. The R Project for Statistical Computing. Available: http://www.r-project.org/ (2003).

32. United States Department of Health and Human Services, Public Health Service, Agency for Toxic Substances and Disease Registry. Toxicological Profile for 4,4'-DDT, 4,4'-DDE, 4,4'-DDD. TP-93/05 (1994).

33. United States Department of Health and Human Services, Public Health Service, Agency for Toxic Substances and Disease Registry. Toxicological Profile for Selected PCBs. TP-92/16 (1993).

34. United States Department of Health and Human Services, Public Health Service, Agency for Toxic Substances and Disease Registry. Toxicological Profile for Chlordane. TP-93/03 (1994).

35. United States Department of Health and Human Services, Public Health Service, Agency for Toxic Substances and Disease Registry. Toxicological Profile for Alpha-, Beta-, Gamma- and Delta-Hexachlorocyclohexane. TP-93/09 (1994).

36. Longnecker MP, Rogan WJ, Lucier G. The human health effects of DDT (dichlorodiphenyltrichloroethane) and PCBS (polychlorinated biphenyls) and an overview of organochlorines in public health. Annu Rev Public Health 18: 211-244 (1997).

37. Patterson DG, Jr., Todd GD, Turner WE, Maggio V, Alexander LR, Needham LL. Levels of non-ortho-substituted (coplanar), mono- and di-ortho-substituted polychlorinated biphenyls, dibenzo-p-dioxins, and dibenzofurans in human serum and adipose tissue. Environ Health Perspect 102 Suppl 1: 195-204 (1994).

38. Woodruff T, Wolff MS, Davis DL, Hayward D. Organochlorine exposure estimation in the study of cancer etiology. Environ Res 65: 132-144 (1994).

39. Feil VJ, Lamoureux CJ, Styrvoky E, Zaylskie RG, Thacker EJ, Holman GM. Metabolism of o,p'-DDT in rats. J Agric Food Chem 21: 1072-1078 (1973).

40. Harrison K. DDT a Banned Insecticide. Available: http://www.chem.ox.ac.uk/mom/ddt/ddt.html (2002).

41. Cocco P, Blair A, Congia P, Saba G, Ecca AR, Palmas C. Long-term health effects of the occupational exposure to DDT. A preliminary report. Ann N Y Acad Sci 837: 246-256 (1997).

42. Dua VK, Pant CS, Sharma VP, Pathak GK. Determination of HCH and DDT in finger-prick whole blood dried on filter paper and its field application for

monitoring concentrations in blood. Bull Environ Contam Toxicol 56: 50-57 (1996).

43. United States Environmental Protection Agency. DDT Regulatory History: A Brief Survey (to 1975). EPA-540/1-75-022 Washington, DC (1975).

44. Waliszewski SM, Aguirre AA, Benitez A, Infanzon RM, Infanzon R, Rivera J. Organochlorine pesticide residues in human blood serum of inhabitants of Veracruz, Mexico. Bull Environ Contam Toxicol 62: 397-402 (1999).

45. Rivero-Rodriguez L, Borja-Aburto VH, Santos-Burgoa C, Waliszewskiy S, Rios C, Cruz V. Exposure assessment for workers applying DDT to control malaria in Veracruz, Mexico. Environ Health Perspect 105: 98-101 (1997).

46. United States Environmental Protection Agency. Technical Factsheet on: Polychlorinated Biphenyls (PCBs). Available: http://www.epa.gov/safewater/dwh/t-soc/pcbs.html (2003).

47. Holoubek I. Polychlorinated Byphenyls (PCBs) World-Wide Contaminated Sites. Brno, Czech Republic: RECETOX - TOCOEN & Associates (2000).

48. Casarett LJ, Doull J. Casarett & Doull's Toxicology: The Basic Science of Poisons. New York: McGraw-Hill (1996).

49. Zook C. 1,1,1-Trichloro-2,2-bis-(4'-chlorophenyl)ethane (DDT) Pathway Map. Available: http://umbbd.ahc.umn.edu/ddt/ddt_map.html (2002).

50. Ruus A, Sandvik M, Ugland KI, Skaare JU. Factors influencing activities of biotransformation enzymes, concentrations and compositional patterns of organochlorine contaminants in members of a marine food web. Aquat Toxicol 61: 73-87 (2002).

51. Donaldson GM, Shutt JL, Hunter P. Organochlorine contamination in bald eagle eggs and nestlings from the Canadian Great Lakes. Arch Environ Contam Toxicol 36: 70-80 (1999).

52. Muir D, Savinova T, Savinov V, Alexeeva L, Potelov V, Svetochev V. Bioaccumulation of PCBs and chlorinated pesticides in seals, fishes and invertebrates from the White Sea, Russia. Sci Total Environ 306: 111-131 (2003).

53. Lagueux J, Pereg D, Ayotte P, Dewailly E, Poirier GG. Cytochrome P450 CYP1A1 enzyme activity and DNA adducts in placenta of women environmentally exposed to organochlorines. Environ Res 80: 369-382 (1999).

54. Dewailly E, Ayotte P, Bruneau S, Gingras S, Belles-Isles M, Roy R. Susceptibility to infections and immune status in Inuit infants exposed to organochlorines. Environ Health Perspect 108: 205-211 (2000).

55. Butler WJ, Seddon L, McMullen E, Houseman J, Tofflemire K, Corriveau A, Weber JP, Mills C, Smith S, Van Oostdam J. Organochlorine levels in maternal and umbilical cord blood plasma in Arctic Canada. Sci Total Environ 302: 27-52 (2003).

56. Muir D, Braune B, DeMarch B, Norstrom R, Wagemann R, Lockhart L, Hargrave B, Bright D, Addison R, Payne J, Reimer K. Spatial and temporal trends and effects of contaminants in the Canadian Arctic marine ecosystem: a review. Sci Total Environ 230: 83-144 (1999).

57. Hayes WJ. Mortality in 1969 from pesticides, including aerosols. Arch Environ Health 31: 61-72 (1976).

58. Herrero C, Ozalla D, Sala M, Otero R, Santiago-Silva M, Lecha M, To-Figueras J, Deulofeu R, Mascaro JM, Grimalt J, Sunyer J. Urinary porphyrin excretion in a human population highly exposed to hexachlorobenzene. Arch Dermatol 135: 400-404 (1999).

59. Durham WF, Armstrong JF, Quinby GE. DDT and DDE content of complete prepared meals. Arch Environ Health 11: 641-647 (1965).

60. Helzlsouer KJ, Alberg AJ, Huang HY, Hoffman SC, Strickland PT, Brock JW, Burse VW, Needham LL, Bell DA, Lavigne JA, Yager JD, Comstock GW. Serum concentrations of organochlorine compounds and the subsequent development of breast cancer. Cancer Epidemiol Biomarkers Prev 8: 525-532 (1999).

61. Walker KC, Goette MB, Batchelor GS. Pesticide residues in foods: dichlorodiphenyltrichloroethane and dichlorodiphenyldichloroethylene content of prepared meals. J Agric Food Chem 2: 1034-1037 (1954).

62. Baker DB, Loo S, Barker J. Evaluation of human exposure to the heptachlor epoxide contamination of milk in Hawaii. Hawaii Med J 50: 108-12, 118 (1991).

63. Wariishi M, Suzuki Y, Nishiyama K. Chlordane residues in normal human blood. Bull Environ Contam Toxicol 36: 635-643 (1986).

64. Duggan RE, Corneliussen PE. Dietary intake of pesticide chemicals in the United States. 3. June 1968--April 1970. Pestic Monit J 5: 331-341 (1972).

65. Duggan RE, Corneliussen PE, Duggan MB, McMahon BM, Martin RJ. Pesticide residue levels in foods in the United States from July 1, 1969, to June 30, 1976: Summary. J Assoc Off Anal Chem 66: 1534-1535 (1983).

66. Fitzgerald EF, Hwang SA, Brix KA, Bush B, Cook K, Worswick P. Fish PCB concentrations and consumption patterns among Mohawk women at Akwesasne. J Expo Anal Environ Epidemiol 5: 1-19 (1995).

67. Clement J, Okey A. Reproduction in female rats born to DDT-treated parents. Bull Environ Contam Toxicol 12: 373-377 (1974).

68. Waliszewski SM, Aguirre AA, Infanzon RM, Silva CS, Siliceo J. Organochlorine pesticide levels in maternal adipose tissue, maternal blood serum, umbilical blood serum, and milk from inhabitants of Veracruz, Mexico. Arch Environ Contam Toxicol 40: 432-438 (2001).

69. Soliman AS, Wang X, DiGiovanni J, Eissa S, Morad M, Vulimiri S, Mahgoub KG, Johnston DA, Do KA, Seifeldin IA, Boffetta P, Bondy ML. Serum organochlorine levels and history of lactation in Egypt. Environ Res 92: 110-117 (2003).

70. Peters HA, Gocmen A, Cripps DJ, Bryan GT, Dogramaci I. Epidemiology of hexachlorobenzene-induced porphyria in Turkey: clinical and laboratory follow-up after 25 years. Arch Neurol 39: 744-749 (1982).

71. Vrecl M, Jan J, Pogacnik A, Bavdek SV. Transfer of planar and non-planar chlorobiphenyls, 4,4'-DDE and hexachlorobenzene from blood to milk and to suckling infants. Chemosphere 33: 2341-2346 (1996).

72. Ribas-Fito N, Sala M, Cardo E, Mazon C, De Muga ME, Verdu A, Marco E, Grimalt JO, Sunyer J. Association of hexachlorobenzene and other organochlorine compounds with anthropometric measures at birth. Pediatr Res 52: 163-167 (2002).

73. Radomski JL, Astolfi E, Deichmann WB, Rey AA. Blood levels of organochlorine pesticides in Argentina: occupationally and nonoccupationally exposed adults, children and newborn infants. Toxicol Appl Pharmacol 20: 186-193 (1971).

74. Dua VK, Pant CS, Sharma VP. Determination of levels of HCH and DDT in soil, water and whole blood from bioenvironmental and insecticide-sprayed areas of malaria control. Indian J Malariol 33: 7-15 (1996).

75. Lees PS, Corn M, Breysse PN. Evidence for dermal absorption as the major route of body entry during exposure of transformer maintenance and repairmen to PCBs. Am Ind Hyg Assoc J 48: 257-264 (1987).

76. Maroni M, Colombi A, Arbosti G, Cantoni S, Foa V. Occupational exposure to polychlorinated biphenyls in electrical workers. II. Health effects. Br J Ind Med 38: 55-60 (1981).

77. Fischbein A, Thornton J, Wolff MS, Bernstein J, Selifoff IJ. Dermatological findings in capacitor manufacturing workers exposed to dielectric fluids containing polychlorinated biphenyls (PCBs). Arch Environ Health 37: 69-74 (1982).

78. Bertazzi PA, Riboldi L, Pesatori A, Radice L, Zocchetti C. Cancer mortality of capacitor manufacturing workers. Am J Ind Med 11: 165-176 (1987).

79. Acquavella JF, Hanis NM, Nicolich MJ, Phillips SC. Assessment of clinical, metabolic, dietary, and occupational correlations with serum polychlorinated biphenyl levels among employees at an electrical capacitor manufacturing plant. J Occup Med 28: 1177-1180 (1986).

80. Stehr-Green PA. Demographic and seasonal influences on human serum pesticide residue levels. J Toxicol Environ Health 27: 405-421 (1989).

81. Ronis MJ, Walker CH. Species variations in the metabolism of liposoluble organochlorine compounds by hepatic microsomal monooxygenase: comparative kinetics in four vertebrate species. Comp Biochem Physiol C 82: 445-449 (1985).

82. Kitamura S, Shimizu Y, Shiraga Y, Yoshida M, Sugihara K, Ohta S. Reductive metabolism of p,p'-DDT and o,p'-DDT by rat liver cytochrome P450. Drug Metab Dispos 30: 113-118 (2002).

83. Gold B, Leuschen T, Brunk G, Gingell R. Metabolism of a DDT metabolite via a chloroepoxide. Chem Biol Interact 35: 159-176 (1981).

84. North HH, Menzer RE. Metabolism of DDT in human embryonic lung cell cultures. J Agric Food Chem 21: 509-510 (1973).

85. Pelletier C, Despres JP, Tremblay A. Plasma organochlorine concentrations in endurance athletes and obese individuals. Med Sci Sports Exerc 34: 1971-1975 (2002).

86. Pelletier C, Doucet E, Imbeault P, Tremblay A. Associations between weight loss-induced changes in plasma organochlorine concentrations, serum T(3) concentration, and resting metabolic rate. Toxicol Sci 67: 46-51 (2002).

87. Borlakoglu JT, Walker CH. Comparative aspects of congener specific PCB metabolism. Eur J Drug Metab Pharmacokinet 14: 127-131 (1989).

88. Ariyoshi N, Oguri K, Koga N, Yoshimura H, Funae Y. Metabolism of highly persistent PCB congener, 2,4,5,2',4',5'-hexachlorobiphenyl, by human CYP2B6. Biochem Biophys Res Commun 212: 455-460 (1995).

89. Glynn AW, Wolk A, Aune M, Atuma S, Zettermark S, Maehle-Schmid M, Darnerud PO, Becker W, Vessby B, Adami HO. Serum concentrations of organochlorines in men: a search for markers of exposure. Sci Total Environ 263: 197-208 (2000).

90. Minh TB, Watanabe M, Tanabe S, Yamada T, Hata J, Watanabe S. Specific accumulation and elimination kinetics of tris(4-chlorophenyl)methane, tris(4-chlorophenyl)methanol, and other persistent organochlorines in humans from Japan. Environ Health Perspect 109: 927-935 (2001).

91. Angerer J, Heinrich R. Occupational exposure to hexachlorocyclohexane. VI. Metabolism of gamma-hexachlorocyclohexane in man. Int Arch Occup Environ Health 52: 59-67 (1983).

92. Mehmood Z, Williamson MP, Kelly DE, Kelly SL. Metabolism of organochlorine pesticides: the role of human cytochrome P450 3A4. Chemosphere 33: 759-769 (1996).

93. Eriksson P, Archer T, Fredriksson A. Altered behaviour in adult mice exposed to a single low dose of DDT and its fatty acid conjugate as neonates. Brain Res 514: 141-142 (1990).

94. Jonsson HT, Jr., Keil JE, Gaddy RG, Loadholt CB, Hennigar GR, Walker EM, Jr. Prolonged ingestion of commercial DDT and PCB; effects on progesterone levels and reproduction in the mature female rat. Arch Environ Contam Toxicol 3: 479-490 (1975).

95. Fabro S, McLachlan JA, Dames NM. Chemical exposure of embryos during the preimplantation stages of pregnancy: mortality rate and intrauterine development. Am J Obstet Gynecol 148: 929-938 (1984).

96. Fitzgerald EF, Standfast SJ, Youngblood LG, Melius JM, Janerich DT. Assessing the health effects of potential exposure to PCBs, dioxins, and furans from electrical transformer fires: the Binghamton State Office Building medical surveillance program. Arch Environ Health 41: 368-376 (1986).

97. Parkinson A, Safe SH, Robertson LW, Thomas PE, Ryan DE, Reik LM, Levin W. Immunochemical quantitation of cytochrome P-450 isozymes and epoxide hydrolase in liver microsomes from polychlorinated or polybrominated biphenyl-treated rats. A study of structure-activity relationships. J Biol Chem 258: 5967-5976 (1983).

98. Frigo DE, Burow ME, Mitchell KA, Chiang TC, McLachlan JA. DDT and its metabolites alter gene expression in human uterine cell lines through estrogen receptor-independent mechanisms. Environ Health Perspect 110: 1239-1245 (2002).

99. Lundberg C. Effect of DDT on cytochrome P-450 and estrogren-dependent functions in mice. Environ Physiol Biochem 4: 200-204 (1974).

100. Korrick SA, Chen C, Damokosh AI, Ni J, Liu X, Cho SI, Altshul L, Ryan L, Xu X. Association of DDT with spontaneous abortion: a case-control study. Ann Epidemiol 11: 491-496 (2001).

101. Hauser R, Altshul L, Chen Z, Ryan L, Overstreet J, Schiff I, Christiani DC. Environmental organochlorines and semen quality: results of a pilot study. Environ Health Perspect 110: 229-233 (2002).

102. Glynn AW, Michaelsson K, Lind PM, Wolk A, Aune M, Atuma S, Darnerud PO, Mallmin H. Organochlorines and bone mineral density in Swedish men from the general population. Osteoporos Int 11: 1036-1042 (2000).

103. Cooper GS, Savitz DA, Millikan R, Chiu KT. Organochlorine exposure and age at natural menopause. Epidemiology 13: 729-733 (2002).

104. Wade MG, Parent S, Finnson KW, Foster W, Younglai E, McMahon A, Cyr DG, Hughes C. Thyroid toxicity due to subchronic exposure to a complex mixture of 16 organochlorines, lead, and cadmium. Toxicol Sci 67: 207-218 (2002).

105. Stehr-Green PA, Wohlleb JC, Royce W, Head SL. An evaluation of serum pesticide residue levels and liver function in persons exposed to dairy products contaminated with heptachlor. JAMA 259: 374-377 (1988).

106. United States Environmental Protection Agency. Health Effects of PCBs. Available: http://www.epa.gov/opptintr/pcb/effects.htm (2002).

107. Richthoff J, Rylander L, Jonsson BA, Akesson H, Hagmar L, Nilsson-Ehle P, Stridsberg M, Giwercman A. Serum levels of 2,2',4,4',5,5'-hexachlorobiphenyl (CB-153) in relation to markers of reproductive function in young males from the general Swedish population. Environ Health Perspect 111: 409-413 (2003).

108. Bloom MS, Weiner JM, Vena JE, Beehler GP. Exploring associations between serum levels of select organochlorines and thyroxine in a sample of New York

state sportsmen: the New York State Angler Cohort Study. Environ Res 93: 52-66 (2003).

109. Belles-Isles M, Ayotte P, Dewailly E, Weber JP, Roy R. Cord blood lymphocyte functions in newborns from a remote maritime population exposed to organochlorines and methylmercury. J Toxicol Environ Health A 65: 165-182 (2002).

110. LeBel G, Dodin S, Ayotte P, Marcoux S, Ferron LA, Dewailly E. Organochlorine exposure and the risk of endometriosis. Fertil Steril 69: 221-228 (1998).

111. Wolff MS, Berkowitz GS, Brower S, Senie R, Bleiweiss IJ, Tartter P, Pace B, Roy N, Wallenstein S, Weston A. Organochlorine exposures and breast cancer risk in New York City women. Environ Res 84: 151-161 (2000).

112. Weiderpass E, Adami HO, Baron JA, Wicklund-Glynn A, Aune M, Atuma S, Persson I. Organochlorines and endometrial cancer risk. Cancer Epidemiol Biomarkers Prev 9: 487-493 (2000).

113. Ward EM, Schulte P, Grajewski B, Andersen A, Patterson DG, Jr., Turner W, Jellum E, Deddens JA, Friedland J, Roeleveld N, Waters M, Butler MA, DiPietro E, Needham LL. Serum organochlorine levels and breast cancer: a nested case-control study of Norwegian women. Cancer Epidemiol Biomarkers Prev 9: 1357-1367 (2000).

114. Schecter A, Toniolo P, Dai LC, Thuy LT, Wolff MS. Blood levels of DDT and breast cancer risk among women living in the north of Vietnam. Arch Environ Contam Toxicol 33: 453-456 (1997).

115. Romieu I, Hernandez-Avila M, Lazcano-Ponce E, Weber JP, Dewailly E. Breast cancer, lactation history, and serum organochlorines. Am J Epidemiol 152: 363-370 (2000).

116. Mendonca GA, Eluf-Neto J, Andrada-Serpa MJ, Carmo PA, Barreto HH, Inomata ON, Kussumi TA. Organochlorines and breast cancer: a case-control study in Brazil. Int J Cancer 83: 596-600 (1999).

117. Porta M, Malats N, Jariod M, Grimalt JO, Rifa J, Carrato A, Guarner L, Salas A, Santiago-Silva M, Corominas JM, Andreu M, Real FX. Serum concentrations of organochlorine compounds and K-ras mutations in exocrine pancreatic cancer. PANKRAS II Study Group. Lancet 354: 2125-2129 (1999).

118. Nordstrom M, Hardell L, Lindstrom G, Wingfors H, Hardell K, Linde A. Concentrations of organochlorines related to titers to Epstein-Barr virus early antigen IgG as risk factors for hairy cell leukemia. Environ Health Perspect 108: 441-445 (2000).

119. Millikan R, DeVoto E, Duell EJ, Tse CK, Savitz DA, Beach J, Edmiston S, Jackson S, Newman B. Dichlorodiphenyldichloroethene, polychlorinated biphenyls, and breast cancer among African-American and white women in North Carolina. Cancer Epidemiol Biomarkers Prev 9: 1233-1240 (2000).

120. Laden F, Hankinson SE, Wolff MS, Colditz GA, Willett WC, Speizer FE, Hunter DJ. Plasma organochlorine levels and the risk of breast cancer: an extended follow-up in the Nurses' Health Study. Int J Cancer 91: 568-574 (2001).

121. Gammon MD, Wolff MS, Neugut AI, Eng SM, Teitelbaum SL, Britton JA, Terry MB, Levin B, Stellman SD, Kabat GC, Hatch M, Senie R, Berkowitz G, Bradlow

HL, Garbowski G, Maffeo C, Montalvan P, Kemeny M, Citron M, Schnabel F, Schuss A, Hajdu S, Vinceguerra V, Niguidula N, Ireland K, Santella RM. Environmental toxins and breast cancer on Long Island. II. Organochlorine compound levels in blood. Cancer Epidemiol Biomarkers Prev 11: 686-697 (2002).

122. Demers A, Ayotte P, Brisson J, Dodin S, Robert J, Dewailly E. Risk and aggressiveness of breast cancer in relation to plasma organochlorine concentrations. Cancer Epidemiol Biomarkers Prev 9: 161-166 (2000).

123. Charlier C, Albert A, Herman P, Hamoir E, Gaspard U, Meurisse M, Plomteux G. Breast cancer and serum organochlorine residues. Occup Environ Med 60: 348-351 (2003).

124. Charles MJ, Schell MJ, Willman E, Gross HB, Lin Y, Sonnenberg S, Graham ML. Organochlorines and 8-hydroxy-2'-deoxyguanosine (8-OHdG) in cancerous and noncancerous breast tissue: do the data support the hypothesis that oxidative DNA damage caused by organochlorines affects breast cancer? Arch Environ Contam Toxicol 41: 386-395 (2001).

125. Dorgan JF, Brock JW, Rothman N, Needham LL, Miller R, Stephenson HE, Jr., Schussler N, Taylor PR. Serum organochlorine pesticides and PCBs and breast cancer risk: results from a prospective analysis (USA). Cancer Causes Control 10: 1-11 (1999).

126. Sturgeon SR, Brock JW, Potischman N, Needham LL, Rothman N, Brinton LA, Hoover RN. Serum concentrations of organochlorine compounds and endometrial cancer risk (United States). Cancer Causes Control 9: 417-424 (1998).

127. Wolff MS, Toniolo PG. Environmental organochlorine exposure as a potential etiologic factor in breast cancer. Environ Health Perspect 103 Suppl 7: 141-145 (1995).

128. Hunter DJ, Hankinson SE, Laden F, Colditz GA, Manson JE, Willett WC, Speizer FE, Wolff MS. Plasma organochlorine levels and the risk of breast cancer. N Engl J Med 337: 1253-1258 (1997).

129. Hoppin JA, Tolbert PE, Holly EA, Brock JW, Korrick SA, Altshul LM, Zhang RH, Bracci PM, Burse VW, Needham LL. Pancreatic cancer and serum organochlorine levels. Cancer Epidemiol Biomarkers Prev 9: 199-205 (2000).

130. Krieger N, Wolff MS, Hiatt RA, Rivera M, Vogelman J, Orentreich N. Breast cancer and serum organochlorines: a prospective study among white, black, and Asian women. J Natl Cancer Inst 86: 589-599 (1994).

131. Rothman N, Cantor KP, Blair A, Bush D, Brock JW, Helzlsouer K, Zahm SH, Needham LL, Pearson GR, Hoover RN, Comstock GW, Strickland PT. A nested case-control study of non-Hodgkin lymphoma and serum organochlorine residues. Lancet 350: 240-244 (1997).

132. Calle EE, Frumkin H, Henley SJ, Savitz DA, Thun MJ. Organochlorines and breast cancer risk. CA Cancer J Clin 52: 301-309 (2002).

133. von Muhlendahl KE. Hormonally active organochlorines and breast cancer: don't believe every abstract. Eur J Pediatr 158: 603-604 (1999).

134. Porta M. Role of organochlorine compounds in the etiology of pancreatic cancer: a proposal to develop methodological standards. Epidemiology 12: 272-276 (2001).

135. Laden F, Collman G, Iwamoto K, Alberg AJ, Berkowitz GS, Freudenheim JL, Hankinson SE, Helzlsouer KJ, Holford TR, Huang HY, Moysich KB, Tessari JD, Wolff MS, Zheng T, Hunter DJ. 1,1-Dichloro-2,2-bis(p-chlorophenyl)ethylene and polychlorinated biphenyls and breast cancer: combined analysis of five U.S. studies. J Natl Cancer Inst 93: 768-776 (2001).

136. Driedger SM, Eyles J. Organochlorines and breast cancer: the uses of scientific evidence in claimsmaking. Soc Sci Med 52: 1589-1605 (2001).

137. Cabral JR, Hall RK, Rossi L, Bronczyk SA, Shubik P. Effects of long-term intake of DDT on rats. Tumori 68: 11-17 (1982).

138. Grimalt JO, Sunyer J, Moreno V, Amaral OC, Sala M, Rosell A, Anto JM, Albaiges J. Risk excess of soft-tissue sarcoma and thyroid cancer in a community exposed to airborne organochlorinated compound mixtures with a high hexachlorobenzene content. Int J Cancer 56: 200-203 (1994).

139. Aronson KJ, Miller AB, Woolcott CG, Sterns EE, McCready DR, Lickley LA, Fish EB, Hiraki GY, Holloway C, Ross T, Hanna WM, SenGupta SK, Weber JP. Breast adipose tissue concentrations of polychlorinated biphenyls and other organochlorines and breast cancer risk. Cancer Epidemiol Biomarkers Prev 9: 55-63 (2000).

140. United States Food and Drug Administration. Action Levels for Poisonous or Deleterious Substances in Human Food and Animal Feed. Available: http://www.cfsan.fda.gov/~lrd/fdaact.html (2004).

141. Biological markers in environmental health research. Committee on Biological Markers of the National Research Council. Environ Health Perspect 74: 3-9 (1987).

142. Waliszewski SM, Aguirre AA, Infanzon RM, Benitez A, Rivera J. Comparison of organochlorine pesticide levels in adipose tissue and human milk of mothers living in Veracruz, Mexico. Bull Environ Contam Toxicol 62: 685-690 (1999).

143. Waliszewski SM, Aguirre AA, Infanzon RM, Lopez-Carrillo L, Torres-Sanchez L. Comparison of organochlorine pesticide levels in adipose tissue and blood serum from mothers living in Veracruz, Mexico. Bull Environ Contam Toxicol 64: 8-15 (2000).

144. Needham LL, Burse VW, Head SL, Korver MP, McClure PC, Andrews JSJr. Adipose tissue/serum partitioning of chlorinated hydrocarbon pesticides in humans. Chemosphere 20: 975-980 (1990).

145. Brock JW, Burse VW, Ashley DL, Najam AR, Green VE, Korver MP, Powell MK, Hodge CC, Needham LL. An improved analysis for chlorinated pesticides and polychlorinated biphenyls (PCBs) in human and bovine sera using solid-phase extraction. J Anal Toxicol 20: 528-536 (1996).

146. Breysse PN, Lees PSJ. Analysis of Gases and Vapors. In: The Occupational Environment - Its Evaluation and Control (DiNardi SR, ed). Fairfax: American Industrial Hygiene Association, 228-241 (1997).

147. Burse VW, Head SL, Korver MP, McClure PC, Donahue JF, Needham LL. Determination of selected organochlorine pesticides and polychlorinated biphenyls in human serum. J Anal Toxicol 14: 137-142 (1990).

148. Cantor KP, Strickland PT, Brock JW, Bush D, Helzlsouer K, Needham LL, Zahm SH, Comstock GW, Rothman N. Risk of non-Hodgkin's lymphoma and prediagnostic serum organochlorines: beta-hexachlorocyclohexane, chlordane/heptachlor-related compounds, dieldrin, and hexachlorobenzene. Environ Health Perspect 111: 179-183 (2003).

149. Burns JE. Pesticides in people: organochlorine pesticide and polychlorinated biphenyl residues in biopsied human adipose tissue-Texas 1969-1972. Pesticides Monitoring Journal 7: 122-126 (1974).

150. Brandt-Rauf PW, Niman HL. Serum screening for oncogene proteins in workers exposed to PCBs. Br J Ind Med 45: 689-693 (1988).

151. Angerer J, Heinzow B, Reimann DO, Knorz W, Lehnert G. Internal exposure to organic substances in a municipal waste incinerator. Int Arch Occup Environ Health 64: 265-273 (1992).

152. Hoyer AP, Grandjean P, Jorgensen T, Brock JW, Hartvig HB. Organochlorine exposure and risk of breast cancer. Lancet 352: 1816-1820 (1998).

153. Hoyer AP, Jorgensen T, Rank F, Grandjean P. Organochlorine exposures influence on breast cancer risk and survival according to estrogen receptor status: a Danish cohort-nested case-control study. BMC Cancer 1: 8 (2001).

154. Hoyer AP, Jorgensen T, Brock JW, Grandjean P. Organochlorine exposure and breast cancer survival. J Clin Epidemiol 53: 323-330 (2000).

155. Fitzgerald EF, Deres DA, Hwang SA, Bush B, Yang BZ, Tarbell A, Jacobs A. Local fish consumption and serum PCB concentrations among Mohawk men at Akwesasne. Environ Res 80: S97-S103 (1999).

156. Kearney JP, Cole DC, Ferron LA, Weber JP. Blood PCB, p,p'-DDE, and mirex levels in Great Lakes fish and waterfowl consumers in two Ontario communities. Environ Res 80: S138-S149 (1999).

157. Baris D, Kwak LW, Rothman N, Wilson W, Manns A, Tarone RE, Hartge P. Blood levels of organochlorines before and after chemotherapy among non-Hodgkin's lymphoma patients. Cancer Epidemiol Biomarkers Prev 9: 193-197 (2000).

158. Fitzgerald EF, Hwang SA, Deres DA, Bush B, Cook K, Worswick P. The association between local fish consumption and DDE, mirex, and HCB concentrations in the breast milk of Mohawk women at Akwesasne. J Expo Anal Environ Epidemiol 11: 381-388 (2001).

159. Fitzgerald EF, Hwang SA, Bush B, Cook K, Worswick P. Fish consumption and breast milk PCB concentrations among Mohawk women at Akwesasne. Am J Epidemiol 148: 164-172 (1998).

160. Balluz L, Philen R, Ortega L, Rosales C, Brock J, Barr D, Kieszak S. Investigation of systemic lupus erythematosus in Nogales, Arizona. Am J Epidemiol 154: 1029-1036 (2001).

161. Nigam SK, Karnik AB, Majumder SK, Visweswariah K, Raju GS, Bai KM, Lakkad BC, Thakore KN, Chatterjee BB. Serum hexachlorocyclohexane residues in workers engaged at a HCH manufacturing plant. Int Arch Occup Environ Health 57: 315-320 (1986).

162. Veith GD, Kuehl DW, Puglisi FA, Glass GE, Eaton JG. Residues of PCB's and DDT in the western Lake Superior ecosystem. Arch Environ Contam Toxicol 5: 487-499 (1977).

163. Belles-Isles M, Ayotte P, Dewailly E, Weber JP, Roy R. Cord blood lymphocyte functions in newborns from a remote maritime population exposed to organochlorines and methylmercury. J Toxicol Environ Health A 65: 165-182 (2002).

164. Comstock GW, Bush TL, Helzlsouer KJ, Hoffman SC. The Washington County Training Center: an exemplar of public health research in the field. Am J Epidemiol 134: 1023-1029 (1991).

165. Glynn AW, Granath F, Aune M, Atuma S, Darnerud PO, Bjerselius R, Vainio H, Weiderpass E. Organochlorines in Swedish women: determinants of serum concentrations. Environ Health Perspect 111: 349-355 (2003).

166. Sweeney AM, Symanski E, Burau KD, Kim YJ, Humphrey HE, Smithci MA. Changes in serum PBB and PCB levels over time among women of varying ages at exposure. Environ Res 86: 128-139 (2001).

167. James RA, Hertz-Picciotto I, Willman E, Keller JA, Charles MJ. Determinants of serum polychlorinated biphenyls and organochlorine pesticides measured in women from the child health and development study cohort, 1963-1967. Environ Health Perspect 110: 617-624 (2002).

168. Bertram HP, Kemper FH, Muller C. Hexachlorobenzene content in human whole blood and adipose tissue: experiences in environmental specimen banking. IARC Sci Publ 173-182 (1986).

169. Phillips DL, Pirkle JL, Burse VW, Bernert JT, Jr., Henderson LO, Needham LL. Chlorinated hydrocarbon levels in human serum: effects of fasting and feeding. Arch Environ Contam Toxicol 18: 495-500 (1989).

170. Maryland Department of the Environment. Expanded Site Inspection of the Central Chemical - Hagerstown Site - MD 302. 2 (1996).

171. United States Environmental Protection Agency. Central Chemical. Available: http://epa.gov/reg3hwmd/super/MD/central-chem/pad.htm (2004).

172. United States Environmental Protection Agency. Cleanup Process. Available: http://epa.gov/superfund/action/process/sfproces.htm (2004).

173. Holman CD, Bass AJ, Rouse IL, Hobbs MS. Population-based linkage of health records in Western Australia: development of a health services research linked database. Aust N Z J Public Health 23: 453-459 (1999).

174. Pittman J, Andrews H, Struening E. The use of zip coded population data in social area studies of service utilization. Eval Program Plann 9: 309-317 (1986).

175. Parchman ML, Ferrer RL, Blanchard KS. Geography and geographic information systems in family medicine research. Fam Med 34: 132-137 (2002).

176. Krieger N, Chen JT, Waterman PD, Soobader MJ, Subramanian SV, Carson R. Geocoding and monitoring of US socioeconomic inequalities in mortality and cancer incidence: does the choice of area-based measure and geographic level matter?: the Public Health Disparities Geocoding Project. Am J Epidemiol 156: 471-482 (2002).

177. McElroy JA, Remington PL, Trentham-Dietz A, Robert SA, Newcomb PA. Geocoding addresses from a large population-based study: lessons learned. Epidemiology 14: 399-407 (2003).

178. Hurley SE, Saunders TM, Nivas R, Hertz A, Reynolds P. Post office box addresses: a challenge for geographic information system-based studies. Epidemiology 14: 386-391 (2003).

179. Floret N, Mauny F, Challier B, Arveux P, Cahn JY, Viel JF. Dioxin emissions from a solid waste incinerator and risk of non-Hodgkin lymphoma. Epidemiology 14: 392-398 (2003).

180. Krieger N, Waterman P, Lemieux K, Zierler S, Hogan JW. On the wrong side of the tracts? Evaluating the accuracy of geocoding in public health research. Am J Public Health 91: 1114-1116 (2001).

181. MacDorman MF, Gay GA. State initiatives in geocoding vital statistics data. J Public Health Manag Pract 5: 91-93 (1999).

182. Chen FM, Breiman RF, Farley M, Plikaytis B, Deaver K, Cetron MS. Geocoding and linking data from population-based surveillance and the US Census to evaluate the impact of median household income on the epidemiology of invasive Streptococcus pneumoniae infections. Am J Epidemiol 148: 1212-1218 (1998).

183. Barcellos C, Sabroza PC. Socio-environmental determinants of the leptospirosis outbreak of 1996 in western Rio de Janeiro: a geographical approach. Int J Environ Health Res 10: 301-313 (2000).

184. Skelly C, Black W, Hearnden M, Eyles R, Weinstein P. Disease surveillance in rural communities is compromised by address geocoding uncertainty: a case study of campylobacteriosis. Aust J Rural Health 10: 87-93 (2002).

185. Krieger N, Chen JT, Waterman PD, Soobader MJ, Subramanian SV, Carson R. Choosing area based socioeconomic measures to monitor social inequalities in low birth weight and childhood lead poisoning: The Public Health Disparities Geocoding Project (US). J Epidemiol Community Health 57: 186-199 (2003).

186. Krieger N, Waterman PD, Chen JT, Soobader MJ, Subramanian S. Monitoring Socioeconomic Inequalities in Sexually Transmitted Infections, Tuberculosis, and Violence: Geocoding and Choice of Area-Based Socioeconomic Measures-The Public Health Disparities Geocoding Project (US). Public Health Rep 118: 240-260 (2003).

187. Krieger N. Place, space, and health: GIS and epidemiology. Epidemiology 14: 384-385 (2003).

188. Roy F.Weston ISATAT. Trip Report. SAT0300595 (1997).

189. Moysich KB, Ambrosone CB, Mendola P, Kostyniak PJ, Greizerstein HB, Vena JE, Menezes RJ, Swede H, Shields PG, Freudenheim JL. Exposures associated with serum organochlorine levels among postmenopausal women from western New York State. Am J Ind Med 41: 102-110 (2002).

190. Draper NR, Simith H. Applied Linear Regression Analysis. New York (1998).

191. Curriero FC, Lele S. A composite Likelihood Aproach to Semivariogram Estimation. Journal of Agricultural, Biological, and Environmental Statistics 4: 9-28 (1999).

192. Ribeiro Jr., P. J. and Diggle, P. J. geoR: A package for geostatistical analysis. R-NEWS [1] (2001).

193. Ribeiro Jr. PJ, Diggle PJ. geoR: A package for geostatistical data analysis using the R software. Available: http://www.est.ufpr.br/geoR/ (2003).

194. Booz Allen Hamilton I. RCRA, Superfund, and EPCRA Hotline Training Module. EPA540-R-98-029 (1998).

195. Hwang SA, Fitzgerald EF, Cayo M, Yang BZ, Tarbell A, Jacobs A. Assessing environmental exposure to PCBs among Mohawks at Akwesasne through the use of geostatistical methods. Environ Res 80: S189-S199 (1999).

196. Preat B. Application of geostatistical methods for estimation of the dispersion variance of occupational exposures. Am Ind Hyg Assoc J 48: 877-884 (1987).

197. Cressie N. Fitting Variogram models by weighted least squares. Journal of the International Association for Mathematical Geology 17: 563-586 (1985).

198. Diggle PJ, Tawn JA, Moyeed RA. Model-based geostatistics. Applied Statistics 47: 299-326 (1998).

199. Myers DE, Begovich CL, Butz TR, Kane VE. Variogram Models for Regional Groundwater Geochemical Data. Mathematical Geology 14: 629-644 (1982).

200. Webster R, Oliver MA. Sample Adequately to Estimate Variograms of Soil Properties. Journal of Soil Science 43: 177-192 (1992).

201. Ormsby T, Napoleon E, Burke R, Groessl C, Feaster L. Getting to Know ArcGIS desktop. Redlands: Environmental Systems Research Institute (2001).

202. Haas TC. Lognormal and Moving Window Methods of Estimating Acid Deposition. Journal of the American Statistical Association 85: 950-963 (1990).

203. Armstrong M, Boufassa A. Comparing the Robustness of Ordinary Kriging and Lognormal Kriging: Outlier Resistance. Mathematical Geology 20: 447-457 (1988).

204. Dowd PA. Lognormal Kriging -- The General Case. Journal of the International Association for Mathematical Geology 14: 475-499 (1982).

205. United States Environmental Protection Agency. Soil Screening Users Guide Technical Document. EPA/540/R-95/128 Region 9 (1996).

206. United States Environmental Protection Agency R3. Risk-Based Concentration Table (2003).

207. Christensen OF, Diggle PJ, Ribeiro Jr. PJ. Analysing positive-valued spatial data: the transformed Gaussian model. In: GeoENVIII - Geostatistics for environmental applications, Vol 11 (Monestiez P, Allard D, Froidevaux R, eds). New York: Kluwer Series, 287-298 (2001).

208. Christensen, O. F. and Ribeiro Jr, P. J. geoRglm: Software for generalised linear spatial models using R. [0.7-5] (2003).

209. Maryland Department of the Environment. Expanded Site Inspection of the Central Chemical - Hagerstown Site - MD 302. 2 (1996).

210. Schildkraut JM, Demark-Wahnefried W, DeVoto E, Hughes C, Laseter JL, Newman B. Environmental contaminants and body fat distribution. Cancer Epidemiol Biomarkers Prev 8: 179-183 (1999).

211. Dallaire F, Dewailly E, Laliberte C, Muckle G, Ayotte P. Temporal trends of organochlorine concentrations in umbilical cord blood of newborns from the

lower north shore of the St. Lawrence river (Quebec, Canada). Environ Health Perspect 110: 835-838 (2002).

212. Waliszewski SM, Pardio VT, Chantiri JN, Infanzon RM, Rivera J. Organochlorine pesticide residues in adipose tissue of Mexicans. Sci Total Environ 181: 125-131 (1996).

213. Boscoe FP, Kielb CL, Schymura MJ, Bolani TM. Assessing and Improving Census Tract Completeness. Journal of Registry Managment 29: 117-120 (2002).

214. Longnecker MP, Ryan JJ, Gladen BC, Schecter AJ. Correlations among human plasma levels of dioxin-like compounds and polychlorinated biphenyls (PCBs) and implications for epidemiologic studies. Arch Environ Health 55: 195-200 (2000).

215. Hornung RW, Reed LD. Estimation of Average Concentration in the Presence of Nondetectable Values. Applied Occupational Environmental Hygiene 5: 46-51 (1990).

216. Sakamoto Y, Ishiguro M, Kitagawa G. Akaike Information Criterion Statistics. D. Reidel Publishing Company (1986).

217. Diggle PJ. Time Series: A Biostatistical Introduction. Oxford University Press (2000).

218. Diggle PJ, Liang KY, Zeger SL. Analysis of Longitudinal Data. Oxford University Press (1998).

219. Carroll RJ, Wu CFJ, Ruppert D. The effect of estimating weighted least squares. Journal of the American Statistical Association 83: 1045-1054 (1988).

220. Neter J, Wasserman W, Kurner MH. Applied Linear Regression Models. Richard D. Irwin, Inc. (1989).

221. Stehr-Green PA, Welty E, Steele G, Steinberg K. Evaluation of potential health effects associated with serum polychlorinated biphenyl levels. Environ Health Perspect 70: 255-259 (1986).

222. Pereg D, Dewailly E, Poirier GG, Ayotte P. Environmental exposure to polychlorinated biphenyls and placental CYP1A1 activity in Inuit women from northern Quebec. Environ Health Perspect 110: 607-612 (2002).

223. Wolff MS, Anderson HA. Correspondence re: J. M. Schildkraut et al., Environmental contaminants and body fat distribution. Cancer Epidemiol. Biomark. Prev., 8: 179-183, 1999. Cancer Epidemiol Biomarkers Prev 8: 951-952 (1999).

224. Bertram HP, Kemper FH, Muller C. Hexachlorobenzene content in human whole blood and adipose tissue: experiences in environmental specimen banking. IARC Sci Publ 173-182 (1986).

225. Hernandez-Valero MA, Bondy ML, Spitz MR, Zahm SH. Evaluation of Mexican American migrant farmworker work practices and organochlorine pesticide metabolites. Am J Ind Med 40: 554-560 (2001).

226. Bonner MR, Han D, Nie J, Rogerson P, Vena JE, Freudenheim JL. Positional accuracy of geocoded addresses in epidemiologic research. Epidemiology 14: 408-412 (2003).

227. Agency for Toxic Substances and Disease Registry. ToxFAQs for Aldrin/Dieldrin. Available: http://www.atsdr.cdc.gov/tfacts1.html (2003).

228. Gotway CA. The Use of Conditional Simulation in Nuclear-Waste-Site Performance Assessment. Technometrics 36: 129-141 (1994).

229. Howe HL. Geocoding NY State Cancer Registry. Am J Public Health 76: 1459-1460 (1986).

230. Krieger N, Waterman P, Chen JT, Soobader MJ, Subramanian SV, Carson R. Zip code caveat: bias due to spatiotemporal mismatches between zip codes and US census-defined geographic areas--the Public Health Disparities Geocoding Project. Am J Public Health 92: 1100-1102 (2002).

APPENDIX A

ADDITIONAL EXPLORATORY MAPS

Appendix A: Additional Exploratory Maps

The following provides a sample of the exploratory maps made with geographic information systems (GIS) and used to detect spatial patterns in the blood organochlorine data. The size of each dot on the map corresponds to the relative level of the organochlorine in the blood of the study participant living at that location within Washington County, Maryland. GIS also allows the user to zoom in on areas of high population density, which was done to take a closer look at spatial patterns.

Figure A.1: Exploratory map of blood DDE levels in Washington County, Maryland, 1974

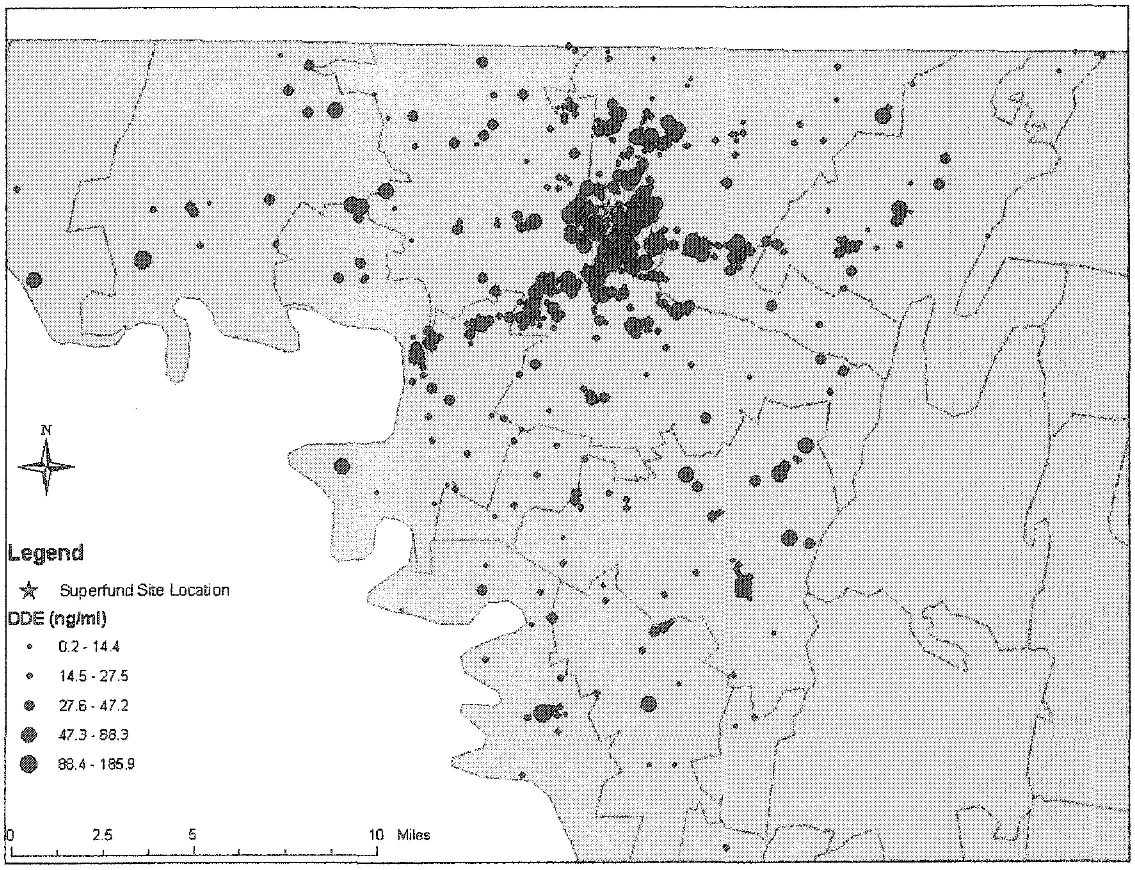


Figure A.2: Exploratory map of blood total PCB levels in Washington County, Maryland, 1974

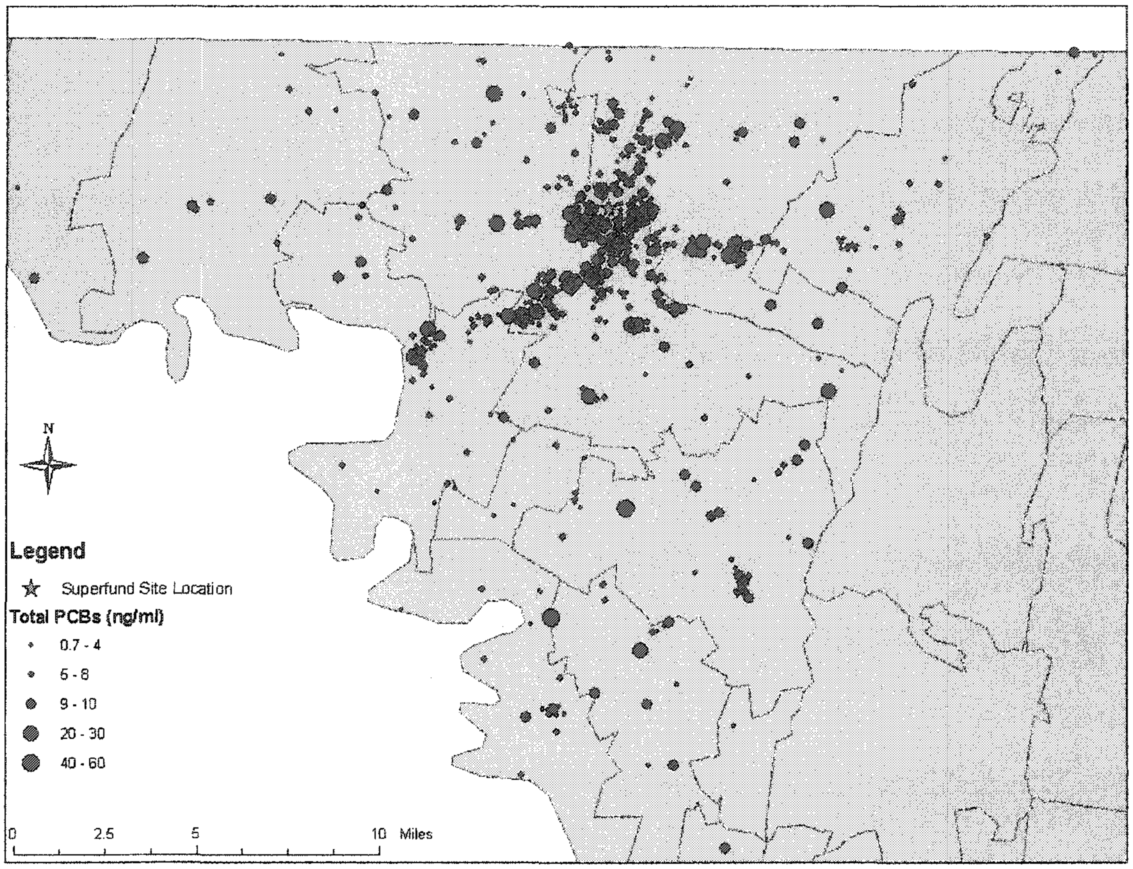


Figure A.3: Exploratory map of lipid adjusted blood dieldrin levels in Washington County, Maryland, 1974

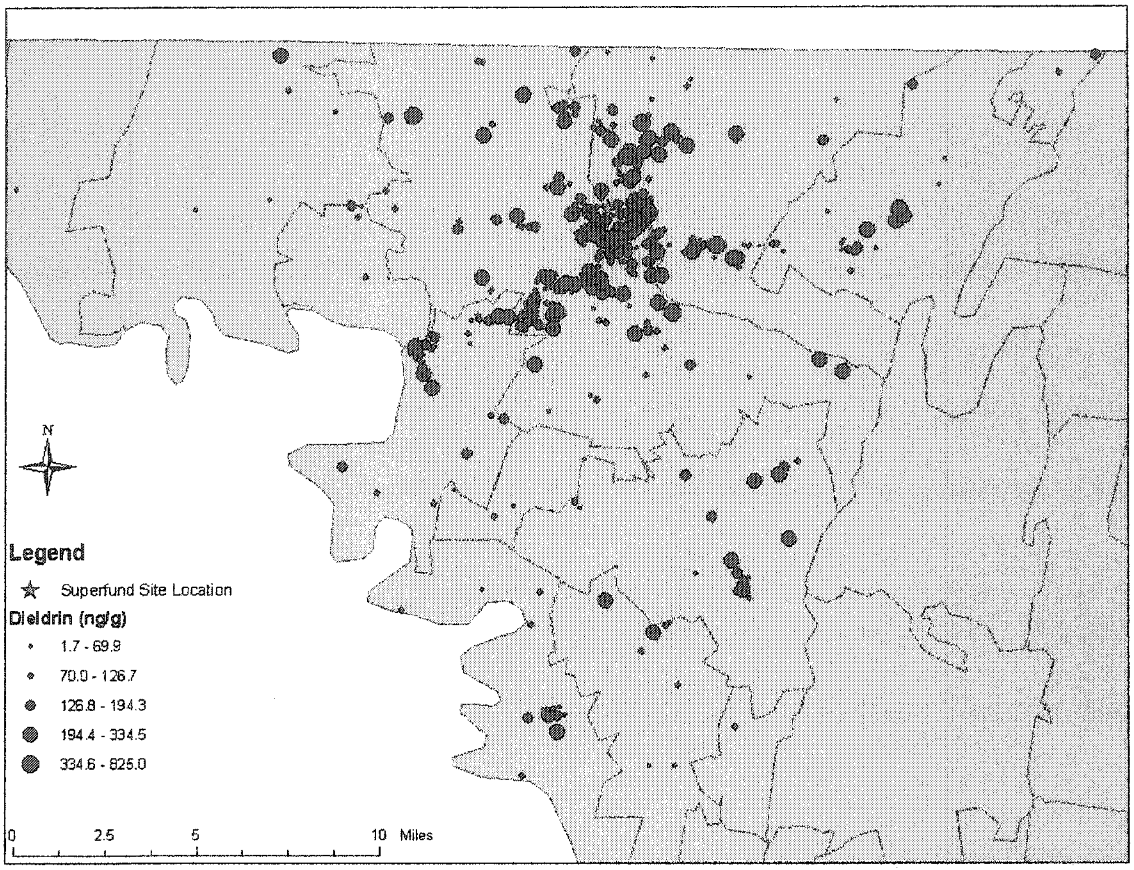


Figure A.4: Exploratory map of blood DDE levels in Washington County, Maryland, 1989

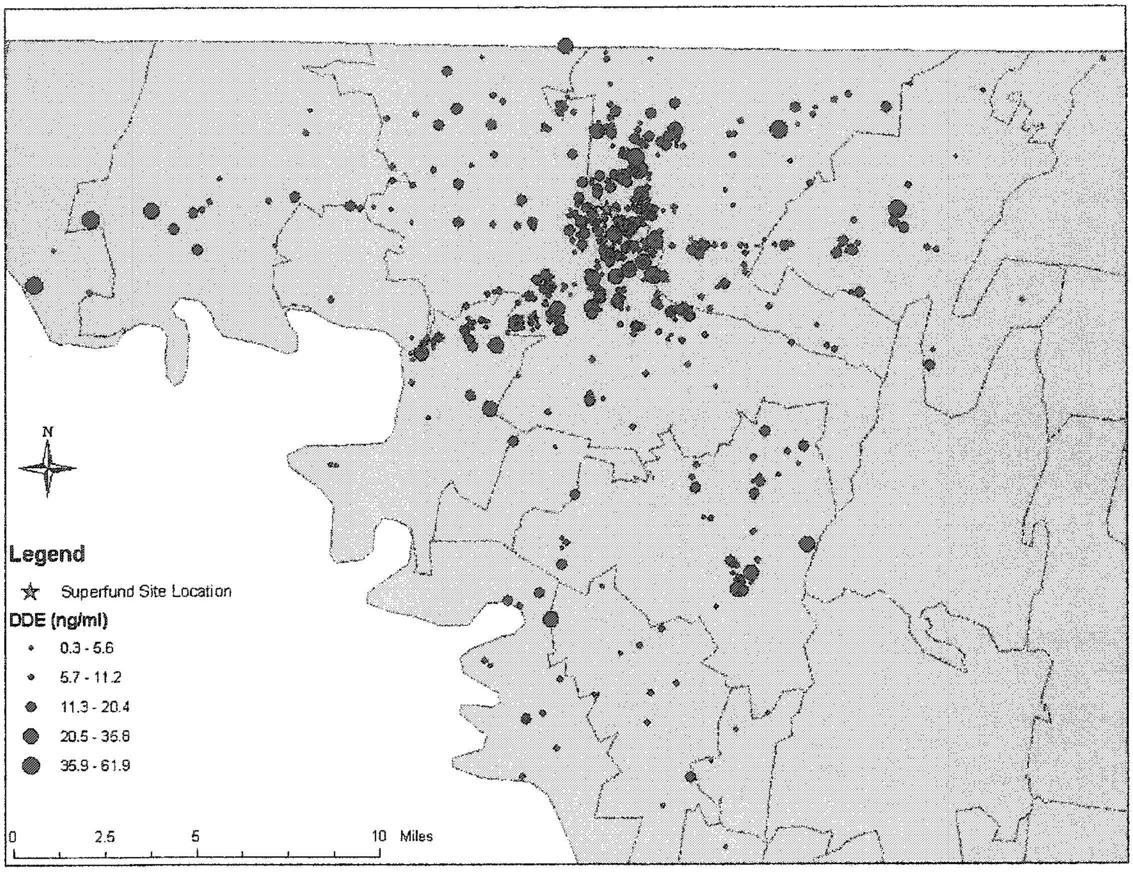
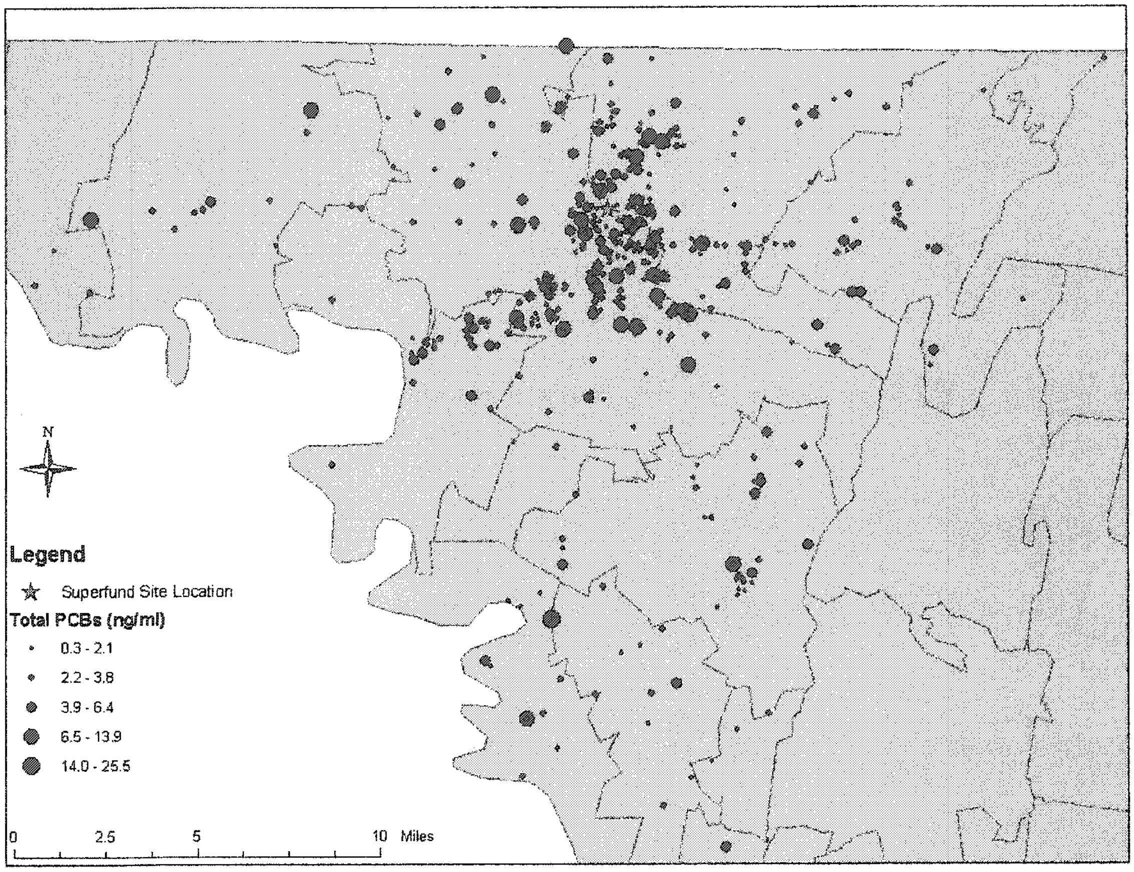


Figure A.5: Exploratory map of blood total PCB levels in Washington County, Maryland, 1989



APPENDIX B

RACIAL EFFECT ON BLOOD DDE

Appendix B: Racial Effect on Blood DDE

It is interesting to note that when both whites and African Americans are considered in the 1974 1,1-dichloro-2,2-*bis*(*p*-chlorophenyl)ethylene (DDE) models discussed in the second manuscript entitled, "The Use of Spatial Information in Determining Potential Risk Factors of Blood Organochlorine Levels in a Population Living Near a Potential Source," race strongly modifies the effect of distance to the site on blood DDE levels. Although distance to the site is not a significant predictor in either African Americans or whites alone, there is a statistically significant difference in the effect of distance to the site on blood DDE levels. However, this finding is only based on a population of 24 African Americans (compared to 1,350 whites), and therefore, the validity of this finding is questionable. However, an attempt is made here to explore the issue further.

When modeling the 1974 blood DDE data, as discussed in the second manuscript, a strongly significant interaction term between race and distance to the Superfund site becomes apparent. Table B.1 contains results of the DDE models including both African Americans and whites and adjusting for race. The interaction between race and distance to the Superfund site is not only significant, but when taken out of the model, the negative significant coefficient of the distance to the site term becomes positive. This suggests a confounding effect. African Americans do have higher mean values of blood DDE than whites (4018 vs. 3004 ng/g) and live closer to the Superfund site (mean distance of 0.93 vs. 4.03 miles from the site) in this study. However, when stratified by race, a very weak association between DDE in the blood and distance to the site is found with the white population (15.53, p-value = 0.32), and a stronger association is found with African Americans (2540, p-value = 0.07). Hence, race is not a confounding factor, but a strong effect modifier.

Table B.1: Results of Generalized Least Squares Regression Determining the Effects of Covariates on Blood DDE Levels Including African Americans, 1974: Parameter Coefficients (p-values)

Covariate	DDE (ng/ml)		DDE (lipid adjusted) (ng/g)	
	Crude	Adjusted	Crude	Adjusted
Distance from CCC site to Residence (miles)	0.14 (0.5)	-13.1 (0.06)	14.1 (0.6)	-2739 (0.008)
Age (years)	0.17 (<0.001)	0.13 (<0.001)	15.7 (0.01)	11.5 (0.045)
Gender (male vs. female)	11.2 (<0.001)	10.8 (<0.001)	1735.7 (<0.001)	1707 (<0.001)
Race (African American vs. White)	5.7 (0.15)	88.1 (0.042)	1126.2 (0.039)	20788 (0.001)
SES (district average based on education and housing index)	-0.5 (0.4)	10.9 (0.049)	-75.7 (0.45)	2594 (0.002)
Smoking Status (non-smoker vs. smoker)	-2.3 (0.023)	-2.4 (0.011)	-291.1 (0.06)	-281 (0.05)
Drinking water Source (spring vs. municipal)	16.9 (<0.001)	15.5 (<0.001)	2089.5 (0.005)	1928 (0.004)
Drinking water Source (well vs. municipal)	3.0 (0.039)	2.9 (0.04)	406.2 (0.065)	410 (0.05)
Distance to site and race Interaction	0.16 (0.45)	13.2 (0.06)	16.7 (0.5)	2751 (0.008)
SES and race interaction	0.31 (0.4)	-10.8 (0.05)	62.6 (0.3)	-2552 (0.002)
Adjusted R squared		0.173		0.1793
Degrees of Freedom		1115		1045

NA: Not in final model

In order to figure out why there is such a difference in effect of residential distance from the site and blood DDE levels, some simple exploratory analyses were conducted (Table B.2). Other than the fact that African Americans live closer to the site, on average, than whites, and also have higher DDE levels in their blood, it is not obvious why there is such a difference between the two populations given these data. African Americans have only slightly lower education and socioeconomic status (SES) levels than whites. Moreover, the mean age of African Americans is younger than whites, which was found in this research to be negatively associated with higher levels of DDE in blood.

Table B.2: Summary Statistics of Potential Risk Factors for Blood Organochlorine Levels Stratified by Race; 1974

	Whites	African Americans
Number	1350	24
Mean Age (years)	53.3	44.8
Mean District Average SES	9.9	8.6
Mean Education (years)	11.3	10.5
Current Smokers	341	15
Men	782	12
City Water Drinkers	900	17
Spring Water Drinkers	10	0
Well Water Drinkers	162	0
Mean DDE (ng/ml)	19.9	23.8
Mean lipid adjusted DDE (ng/g)	3004	4018
Mean Distance to Site (miles)	4.0	0.93

It is not uncommon in the literature for researchers to report a racial effect on levels of organochlorines in blood. For example, James *et al.* reported that non-white women (primarily African American) had on average, 50 percent higher levels of 1,1,1-trichloro-2,2-bis(*p*-chlorophenyl)ethane (DDT) and DDE in their blood serum (*167*). Several other studies have also reported a racial effect on blood organochlorine levels especially between the higher levels in African Americans than whites (*4,41,130,210,221*).

The reasons for a difference between African American and white blood organochlorine levels are unclear in this study. Speculating, possible reasons that African Americans may have higher blood organochlorine levels than whites include: consuming diets higher in fat, living closer to potential sources, holding occupations with potentially greater exposure to these compounds, or possible metabolic differences between the two races. Hence, the racial effect on blood organochlorine levels in this study, and in general, deserves further consideration.

APPENDIX C

SELECTED R STATISTICAL COMPUTING CODE

Appendix C: Selected R Statistical Computing Code

Selected Statistical Code for Manuscript One

The following is a sample of the R statistical Computing Environment code used to analyze the soil DDE data collected from 1993-1997 to produce the results reported in Manuscript One. The geographical data is in Maryland 1983 State Plane Coordinates measured in feet.

```
# 1) Data Management
# First, the data in text format is read into R

soil<-read.table("c:\\shenshaw\\Thesis\\Environmental_Data
\\soilanalysis\\soil3.csv",header=T,sep=",",as.is=TRUE)
names(soil)
Year<-as.numeric(substr(soil$MAP.ID,1,2))
soil<-data.frame(soil,Year=Year)

# Use only the data in the 90's (ie. 92-97)
u<-soil$Year >90
soil14<-soil[u,]

# Use only DDE data
u<-soil14$ANALYTE=="4,4'-DDE" | soil14$ANALYTE=="DDE"
soil14<-soil14[u,]

# Make a table in order to see if there are any values that are non
detects or at the detection limit
table(soil14$Year,soil14$CONCENTRAT)
table(soil14$Year,soil14$Area)

# For those samples that were analyzed in duplicate, the values were
averaged, and the average value was named the first sample ID with an X
at the end, and the 2 duplicates were taken out of the analysis
soil14[soil14$SAMPLE.ID=="slh585"|soil14$SAMPLE.ID=="slh657",]
soil14$SAMPLE.ID[soil14$SAMPLE.ID=="slh585"]<-"slh585x"
soil14$NEW.CONC[soil14$SAMPLE.ID=="slh585x"]<-mean(c(6600,9500))
dim(soil14)

u<-c(1:118)[soil14$SAMPLE.ID=="slh657"]
soil14<-soil14[-u,]
soil14[soil14$SAMPLE.ID=="slh585x",]
soil14[soil14$SAMPLE.ID=="slh1214"|soil14$SAMPLE.ID=="slh3377",]
soil14$SAMPLE.ID[soil14$SAMPLE.ID=="slh1214"]<-"slh1214x"
soil14$NEW.CONC[soil14$SAMPLE.ID=="slh1214x"]<-mean(c(540,4000))
dim(soil14)

u<-c(1:117)[soil14$SAMPLE.ID=="slh3377"]
soil14<-soil14[-u,]
soil14[soil14$SAMPLE.ID=="slh1214x",]
soil14[soil14$SAMPLE.ID=="slh1217"|soil14$SAMPLE.ID=="slh3404",]
soil14$SAMPLE.ID[soil14$SAMPLE.ID=="slh1217"]<-"slh1217x"
soil14$NEW.CONC[soil14$SAMPLE.ID=="slh1217x"]<-mean(c(2900,2400))
dim(soil14)
```

```
u<-c(1:116)[soil14$SAMPLE.ID=="slh3404"]
soil14<-soil14[-u,]
soil14[soil14$SAMPLE.ID=="slh1217x",]
soil14[soil14$SAMPLE.ID=="slh3444"|soil14$SAMPLE.ID=="slh3459",]
soil14$SAMPLE.ID[soil14$SAMPLE.ID=="slh3444"]<-"slh3444x"
soil14$NEW.CONC[soil14$SAMPLE.ID=="slh3444x"]<-mean(c(820,760))
dim(soil14)
u<-c(1:115)[soil14$SAMPLE.ID=="slh3459"]
soil14<-soil14[-u,]
soil14[soil14$SAMPLE.ID=="slh3444x",]

# Samples taken at depths greater than 1 foot were taken out of the
analysis
u<-
soil14$SAMPLE.ID=="slh861"|soil14$SAMPLE.ID=="slh893"|soil14$SAMPLE.ID=
="slh925"|soil14$SAMPLE.ID=="slh957"
soil14<-soil14[!u,]
dim(soil14)

# In order to get 5 random samples from each quartile of data, the
following code is necessary

summary(soil14$NEW.CONC)
n<-length(soil14$NEW.CONC)
u1<-c(1:110)[soil14$NEW.CONC<=2100]
u2<-sample(u1,5)
u3<-c(1:110)[soil14$NEW.CONC>2100&soil14$NEW.CONC<=4400]
length(u3)
u4<-sample(u3,5)
u5<-c(1:110)[soil14$NEW.CONC>4400&soil14$NEW.CONC<=23000]
u6<-sample(u5,5)
u7<-c(1:110)[soil14$NEW.CONC>23000]
u8<-sample(u7,5)
u.delete<-c(u2,u4,u6,u8)

# For Paper 1, the 5 randomly chosen values taken from each quartile of
the data are identified as follows:
u.delete<-
c(94,65,93,72,85,82,11,74,84,71,107,14,53,101,5,104,32,10,27,110)

# Their sample ID's:
soil14$SAMPLE.ID[u.delete]

# The data is divided into areas

table(soil14$Year,soil14$Area)
Area<-soil14$Area

u<-soil14$Area==1
soil14a1<-soil14[u,]
dim(soil14a1)

u<-soil14$Area==2
soil14a2<-soil14[u,]
dim(soil14a2)
```

```
u<-soil14$Area==3
soil14a3<-soil14[u,]
dim(soil14a3)

u<-soil14$Area==4
soil14a4<-soil14[u,]
dim(soil14a4)

u<-soil14$Area==5
soil14a5<-soil14[u,]
dim(soil14a5)

# A geodataset is made of the entire dataset minus the validation
points
soil16<-soil14[-u.delete,]
soil16<-as.geodata(soil16,data.col=9)

# A geodataset is made of just the validation points
soil17<-soil14[u.delete,]
soil17<-as.geodata(soil17,data.col=9)

# A geodataset is made of all the points
soil18<-as.geodata(soil14,data.col=9)

# A geodatset is made of each area, and plotted
soil14a1<-as.geodata(soil14a1,data.col=9)
points(soil14a1)
title("Soil sample Locations 92-97 minus predictions Area 1, size
relative to level of DDE")
sort(soil14a1$data)

soil14a2<-as.geodata(soil14a2,data.col=9)
points(soil14a2)
title("Soil sample Locations 92-97 minus predictions Area 2, size
relative to level of DDE")
sort(soil14a2$data)

soil14a3<-as.geodata(soil14a3,data.col=9)
points(soil14a3)
title("Soil sample Locations 92-97 minus predictions Area 3, size
relative to level of DDE")
sort(soil14a3$data)

soil14a4<-as.geodata(soil14a4,data.col=9)
points(soil14a4)
title("Soil sample Locations 92-97 minus predictions Area 4, size
relative to level of DDE")
sort(soil14a4$data)

soil14a5<-as.geodata(soil14a5,data.col=9)
points(soil14a5)
title("Soil sample Locations 92-97 minus predictions Area 5, size
relative to level of DDE")
sort(soil14a5$data)

# 2) Data Analysis
# Kriging to Predict levels of DDE at unknown locations.  A
semivariogram of the logged data is first plotted to determine if there
is spatial dependence in the data and to identify initial paramaters
used to model the semivariogram in the data

# First, a semivariogram model with the initial parameters (nugget,
sill, and range) specified is fit to the data
t0<-variog(soil16,max.dist=750,lambda=0)
t1<-variofit(t0,ini.cov.pars=c(2,300),nugget=2.2)

# MakePairs is the function used to get the composite liklihood method
to model the semivariogram.  It is defined as:

MakePairs<-function(N){
  i<-rep(c(1:(N-1)),c((N-1):1))
  j<-numeric(0)
  for(k in 2:N)
    j<-c(j,c(k:N))
  return(cbind(i,j))
}

xy<-soil16$coords
z<-log(soil16$data)
N<-length(soil16$data)
Pairs<-MakePairs(N)

# A spherical model is chosen to model the semivariogram. The composite
likelihood code for the function CLSpherical is:

CLSpherical<-function(theta){
  p1<-Pairs[,1]
  p2<-Pairs[,2]
  tau2<-theta[1]
  sigma2<-theta[2]
  phi<-theta[3]
  ds<-sqrt((xy[p1,1]-xy[p2,1])^2+
           (xy[p1,2]-xy[p2,2])^2)
  kap<-0.5
  gamma<-tau2 + sigma2*((3/2)*(ds/phi) - (1/2)*(ds/phi)^3)
  gamma[ds>=phi]<-tau2+sigma2
  lcl<-sum(((z[p1]-z[p2])^2/(2*gamma)) + log(gamma))
  return(lcl)
}

t2<-nlmP(CLSpherical,c(2,2,300),lower=c(0.01,0.01,0.01),
upper=c(Inf,Inf,Inf))
soil16.cl<-t1
soil16.cl$cov.model<-"spherical"
soil16.cl[[1]]<-t2[[2]][1]
soil16.cl[[2]]<-t2[[2]][-1]
lines(soil16.cl,lty=3)

# Global/Internal cross Validation to determine how well model predicts
at known locations within the dataset

# First, using the geodataset of the 90 points (110 points minus the 20
validation points)
```

```
soil16.xvalid<-xvalid(soil16,model=soil16.cl)
soil16.cv3<-sqrt(mean(soil16.xvalid$error^2))
soil16.cv3

# Internal Cross validation by Area
# (Areas of the map were divided up to see if one area had better
prediction rates than other areas).

# Area 1= the fenceline (inside property)
soil16a1.cv3<-sqrt(mean(soil16.xvalid$error[Area==1]^2))

# Area 2= the trench
soil16a2.cv3<-sqrt(mean(soil16.xvalid$error[Area==2]^2))

# Area 3= the bottom left side of the map (outside property)
soil16a3.cv3<-sqrt(mean(soil16.xvalid$error[Area==3]^2))

# Area 4= Brighton manor subdivision (NW of site outside property)
soil16a4.cv3<-sqrt(mean(soil16.xvalid$error[Area==4]^2))

# Area 5= Residential area NE of site
soil16a5.cv3<-sqrt(mean(soil16.xvalid$error[Area==5]^2))

# Now, the model used for the geodataset of 90 points is used to
predict at the 20 validation locations and the results of the
predictions are compared to the actual observed values at that
location. This is called external cross validation

soil16.krige<-
krige.conv(soil16,loc=soil17$coords,krige=krige.control(obj.model=soil1
6.cl,lambda=0))
soil16.cv3.2<-sqrt(mean((soil16.krige$predict - soil17$data)^2))

# A plot of this prediction is produced:
locs.site<-expand.grid(MD83ft.X=seq(1107800,1109200,length=50),
                       MD83ft.Y=seq(724850,726450,length=50))
soil16.krige.site<-krige.conv(soil16,loc=locs.site,
krige=krige.control(obj.model=soil16.cl,lambda=0))
image.kriging(soil16.krige.site)
points(soil16$coords)

# Information for Table 1 is produced
cbind(soil14$Area[u.delete],soil16.krige$predict,soil17$data)

# Analyses using all 110 data points

# First, plot the semivariogram, and estimate the semivariogram model
based on initial parameters:
plot(variog(soil18,max.dist=750,lambda=0),ylim=c(0,5))
lines.variomodel(cov.model="spherical",cov.pars=c(2,400),nugget=2,max.d
ist=750)
title("Estimating a Variogram model for logged DDE data")

# A variogram is fit to the data based on initial parameters
t0<-variog(soil18,max.dist=750,lambda=0)
t1<-variofit(t0,ini.cov.pars=c(2,400),nugget=2)
```

```
# A spherical variogram model is fit to the data using the composite
liklihood method
xy<-soil18$coords
z<-log(soil18$data)
N<-length(soil18$data)
Pairs<-MakePairs(N)

t2<-nlmP(CLSpherical,c(2,2,400),lower=c(0.01,0.01,0.01),
upper=c(Inf,Inf,Inf))
soil18.cl<-t1
soil18.cl$cov.model<-"spherical"
soil18.cl[[1]]<-t2[[2]][1]
soil18.cl[[2]]<-t2[[2]][-1]
lines(soil18.cl,lty=3)

# Graphs for paper 1

# To see what the quartiles in the data look like for the legend in the
histogram
summary(soil18$data/1000)

# To create a histogram of the data (Figure 2a)
hist(soil18$data/1000,xlab="1992-1997 Levels of DDE in Soil Samples
(ppm)",main=" ")
legend(locator(1),legend=c("N = 110","Mean = 25.40","Stdev = 46.38",
                             "Min = 0.005","Max = 300"))
# To change ppb to ppm
soil18a<-soil18
soil18a$data<-soil18a$data/1000

# To create a plot of the soil locations around the buildings, the
building points had to first be read in from a data file
bldg<-read.table("c:\\shenshaw\\Thesis\\GIS\\Geocode\\CCC\\
buildingsft.txt",header=T)
bldg$building.c<-as.character(bldg$building.c)

# A vector file had to be created made out of the points of the corners
of buildings
types<-as.vector(unique(bldg$building.c))
types<-types[-c(1,9,12,13,14)]

# A geodataset was created made out of the data values divided into
quartiles to produce # Figure 2b.  Went into points.geodata and changed
Y to DDE and saved as Xpoints.geodata

# Figure 3 is created by plotting the semivariogram, and plotting a
semivariogram model # overtop
par(pch=19)
plot(variog(soil18a,max.dist=750,lambda=0),ylim=c(0,5),xlab="Distance
(ft)")
lines(soil18.cl,lwd=2)
legend(locator(1),legend=c("Sill = 2.8","Nugget = 1.5","Range =
339"),ncol=1)
par(pch=1)

# Global/Internal cross Validation using all data is used
# to determine the predictibility of the model based on all 110
datapoints

soil18.xvalid<-xvalid(soil18,model=soil18.cl)
soil18.cv3<-sqrt(mean(soil18.xvalid$error^2))

# The model based on the entire dataset is used to krige at all
locations in a grid over the site
locs.site<-expand.grid(MD83ft.X=seq(1107800,1109200,length=50),
                       MD83ft.Y=seq(724850,726450,length=50))
soil18.krige.site<-krige.conv(soil18,loc=locs.site,

krige=krige.control(obj.model=soil18.cl,lambda=0))
image.kriging(soil18.krige.site)
points(soil18$coords)

# Exporting these kriged results so that they can be brought into
ArcGis to Make Figure 4
CLresults2<-cbind(locs.site,soil18.krige.site$predict)
write(t(CLresults2),file="c:\\shenshaw\\Thesis\\Environmental_Data\\soi
lanalysis\\fig4res.txt",ncol=3)

#In GIS you can produce a better looking image of the kriged values
with a better legend and all the other site features, etc.
```

Selected Statistical Code for Manuscript Two

CLUE I (1974) Analysis

The following consists of selected R Statistical Computing Environment code used to analyze the serum organochlorine data as reported in Manuscript Two. However, in an effort to be brief, but still inclusive of all data analysis, only the analysis of one compound, DDE, is shown here. However, these same analyses were also completed for DDT, DDD, all PCB congeners, total PCBs, hexachlorobenze, beta- and gamma-hexachlorocylcohexane, dieldrin, aldrin, mirex, heptachlor, heptachlor epoxide, alpha- and gamma-chlordane both lipid adjusted and unadjusted.

```
###  Reading in all of the data from text files ###

#  Organochlorine Data

# Clue 1: 1974
oc1<-read.csv("c:\\shenshaw\\Thesis\\CLUE\\CLUE_Data\\
Changed_Variable_Names\\Merged_Data\\wcc1ocdNA.csv",header=T,sep=",")

# Clue 1: breast cancer lipid data
bclip<-read.csv("c:\\shenshaw\\Thesis\\CLUE\\CLUE_Data\\ExtraVars
\\bclip.csv",header=T,sep=",")

# Clue 1: prostate cancer lipid data
prlipc1<-read.csv("c:\\shenshaw\\Thesis\\CLUE\\CLUE_Data\\ExtraVars
\\prlipc1.csv",header=T,sep=",")

# Address Data: clue 1: 1974
addr1<-read.csv("c:\\shenshaw\\Thesis\\GIS\\Geocode\\clue1\\
clue1shpNAwoxydup.csv",header=T,sep=",")
dim(addr1)

# Extra Variable data
# Breast Cancer Studies 1995
bc1<-read.csv("c:\\shenshaw\\Thesis\\CLUE\\CLUE_Data\\ExtraVars
\\bcsvars1NA.csv",header=T,sep=",")
bc2<-read.csv("c:\\shenshaw\\Thesis\\CLUE\\CLUE_Data\\ExtraVars
\\bcsvars2NA.csv",header=T,sep=",")
bc3<-read.csv("c:\\shenshaw\\Thesis\\CLUE\\CLUE_Data\\ExtraVars
\\bcsvars3NA.csv",header=T,sep=",")
bc4<-read.csv("c:\\shenshaw\\Thesis\\CLUE\\CLUE_Data\\ExtraVars
\\bcsvars4NA.csv",header=T,sep=",")

# CLUE Studies
# Clue 1 and 75 census (1974-75)
clue1<-read.csv("c:\\shenshaw\\Thesis\\CLUE\\CLUE_Data\\ExtraVars
\\c1vars1NA.csv",header=T,sep=",")

# CLUE Follow up studies 1996
fol96<-
read.csv("c:\\shenshaw\\Thesis\\CLUE\\CLUE_Data\\ExtraVars\\c2med96NA.c
sv",header=T,sep=",")
dim(fol96)
```

```
names(fol96)

# Round 2 Study 1998
rnd2<-read.csv("c:\\shenshaw\\Thesis\\CLUE\\CLUE_Data\\ExtraVars
\\c2rnd2NA.csv",header=T,sep=",")

# 2000
fol00<-read.csv("c:\\shenshaw\\Thesis\\CLUE\\CLUE_Data\\ExtraVars
\\c2med00NA.csv",header=T,sep=",")

# Diet Study: 1989
diet<-read.csv("c:\\shenshaw\\Thesis\\CLUE\\CLUE_Data\\ExtraVars
\\c2diet89NA.csv",header=T,sep=",")

# Additional variables asked for from several studies
mixed<-read.csv("c:\\shenshaw\\Thesis\\CLUE\\CLUE_Data\\ExtraVars
\\mixedNA.csv",header=T,sep=",")

### Data Management and Exploratory Analysis ###

clue1id<-oc1$clue1

# Working with OC data (Clue 1) Outcome variables
# ng/mL and ng/g (lipid adjusted)

# DDE, looking at available data using the table function
# ng/mL and ng/g (lipid adjusted)
table(is.na(oc1$pdde.ml))
table(is.na(oc1$pdde.lip))

# Make total lipids variables by using cholesterol and triglyceride
data in Philips et al short equation..equation 2, then make it a factor
to multiply unadjusted values by to get ng/g units

t<-match(oc1$clue1,bclip$CLUE)
chol<-bclip$TC[t]
tri<-bclip$TG[t]
oc1<-data.frame(oc1,chol=chol,tri=tri)
tt<-match(oc1$clue1,prlipc1$clue1)
choles<-prlipc1$choles[tt]
oc1<-data.frame(oc1,choles=choles)

n<-length(oc1$tl1)
tlip<-rep(NA,n)
oc1<-data.frame(oc1,tlip=tlip)
oc1$tlip[oc1$study==1]<-
1/(2.27*oc1$chol[oc1$study==1]+oc1$tri[oc1$study==1]+0.623)*100000
oc1$tlip[oc1$study==2]<-
1/(2.27*oc1$choles[oc1$study==2]+oc1$trigly[oc1$study==2]+0.623)*100000
oc1$tlip[oc1$study==3]<-
1/(2.27*oc1$cholestr[oc1$study==3]+oc1$triglycr[oc1$study==3]+0.623)*10
0000

# get rid of outliers in the lipid adjuster values (those that
incorrect lipid adjusted information was recorded) a total of 3
oc1$tlip[oc1$tlip>1000]<-NA
```

```
# need to change all lipid adjusted variables so that they are the same

oc1$pdde.lip<-oc1$pdde.ml*oc1$tlip
#Check for zeros in the data, if there are make = LOD/sqrt(2)

# DDE
# To find out where the zeros are:
u<-c(1:1391)[oc1$pdde.ml==0]
zerodde<-length(oc1$pdde.ml[oc1$pdde.ml==0&!is.na(oc1$pdde.ml)])
/length(oc1$pdde.ml[!is.na(oc1$pdde.ml)])*100
# To find out what study the zeros are found in
oc1[u,5]
# Make them the LOD/sqrt(2)
oc1$pdde.ml[oc1$pdde.ml==0]<-0.61/sqrt(2)
zeroddelip<-length(oc1$pdde.lip[oc1$pdde.lip==0&!is.na(oc1$pdde.lip)])
/length(oc1$pdde.lip[!is.na(oc1$pdde.lip)])*100

# Check for correlations in the DDT/DDE/DDD data
cor(oc1$pdde.ml,oc1$oddt.ml, use="pairwise.complete.obs")
cor(oc1$pdde.ml,oc1$pddt.ml, use="pairwise.complete.obs")
cor(oc1$pdde.ml,oc1$pddd.ml, use="pairwise.complete.obs")
cor(oc1$pdde.lip,oc1$oddt.lip, use="pairwise.complete.obs")
cor(oc1$pdde.lip,oc1$pddt.lip, use="pairwise.complete.obs")
cor(oc1$pdde.lip,oc1$pddd.lip, use="pairwise.complete.obs")

#### getting all covariates ####

t1<-match(oc1$clue1,addr1$CLUE1)
t2<-match(oc1$clue1,bc1$Clue)
t3<-match(oc1$clue1,bc2$Clue)
t4<-match(oc1$clue1,clue1$Clue1)
t5<-match(oc1$clue1,bc3$Clue1)
t6<-match(oc1$clue1,mixed$Clue1)

#1974 variables = t1 & t4 & smoke2 from t6
#1995 variables = t2,t3, & t5

### Continuous Variables ###

# Centroid of site in MD state plane coords 83 ft
sitex<-1108776
sitey<-725732

# residence: zip and x and y location in MD state plane coordinates
zip<-addr1$ARC.ZONE[t1]
addX<-addr1$MD83FT.X[t1]
addY<-addr1$MD83FT.Y[t1]

# Distance from residence location to site measured in miles
dtosite<-(sqrt((addX-sitex)^2+(addY-sitey)^2))/5280

# Age in years
age<-clue1$Age[t4]

# Weight w/o shoes in pounds
weight<-bc3$Wgt[t5]
```

```
# Height in inches
height<-bc3$Htinch[t5] + bc3$Htfeet[t5]*12

# Body Mass index
BMI<-(weight/height^2)*703

# district average SES (in private census districts):
# =(average of years of education and housing index)
SES<-clue1$DaveSes[t4]

# Years of education
edu<-clue1$Ed[t4]

# Months of breastfeeding
brfeed<-bc1$Breastfd[t2]

# Total years applied insecticides on pets
pestpets<-bc2$petyrs[t3]

# Total years applied insecticides on farm animals
pestfarm<-bc2$farmyrs[t3]

# Total years applied insecticides outside house
pestout<-bc2$outyrs[t3]

# Total years applied insecticides inside house
pestin<-bc2$inyrs[t3]

# Total years applied pesticides
pestpest<-bc2$pestyrs[t3]

### Binary Variables ###

# Race (1=white,2=black)
race<-clue1$Race[t4]

# Gender (1=male,2=female)
gender<-clue1$Sex[t4]

# Exposed to pesticides,insecticides,herbicides,fungicides,or fumigants
on the job
# (0=no,1=yes)
occ<-bc2$pestcid[t3]

# current smoker at time of blood draw (1=yes,2=no)
smoke<-clue1$Now[t4]

# how many drinks/week subject normally has (0=less than 1,1=1,2=2-
3,3=4-5,4=6-8,5=9-11,6=12-14,7=15-19,8=20-24,9=25-29,10=30-34,11=35-
39,12=40+)
alcohol<-bc1$Numdrk[t2]

# Are you a vegetarian now (avoid meat) (0=no,1=yes)
meat<-bc3$Vegnow[t5]

# Avoid eggs (0=no,1=yes)
eggs<-bc3$Noeggs[t5]
```

```
# Avoid dairy products (0=no,1=yes)
dairy<-bc3$Nodairy[t5]

# Ever been a vegetarian (0=no,1=yes)
veg<-bc1$Vegdiet[t2]

### Categorical variables ###

# smoking recode:(0=never smoked,1=ex cig smoker,2=ex
cigar/pipe,3=current cig 0-14, 4=current cig 15-24,5=current cig
25+,6=current cig, amount unknown,7=current pipe/cigar, 8=others
smoke3<-mixed$Smkr[t6]

#Recode smoking again to make smoke 2 variable (0=never,1=ex
smoker,2=current smoker)

smoke2<-smoke3
smoke2[smoke3==0]<-0
smoke2[smoke3==1|smoke3==2]<-1
smoke2[smoke3==3|smoke3==4|smoke3==5|smoke3==6|smoke3==7|smoke3==8]<-2

# how often eat fish (0=never or <1/month,1=1/month,2=2-
3/month,3=1/week,4=2/week, 5=3-4/week,6=5-6/week,7=1/day,8=2/day
fish<-bc1$Fish[t2]

# Drinking water source: 0=spring,1=city,2=deep well,3=shallow
well,4=cistern city,5=cistern rain,6=cistern well,7=cistern other or
mixed,8=bottled,9=not stated
water2<-clue1$DrW[t4]

# Recode to make city water the reference category:
# 0=city,1=spring,2=well,3=cistern (rain, mixed, or other),4=bottled

water<-water2
water[water2==1|water2==4]<-0
water[water2==0]<-1
water[water2==2|water2==3|water2==6]<-2
water[water2==5|water2==7]<-3
water[water2==8]<-4
water[water2==9]<-NA

# Marital status: 1=never married,2=married
now,3=separated,4=divorced,5=widowed
marital<-clue1$MS[t4]

# years lived at this residence (1974-year moved in)
res<-74-clue1$YMvd[t4]
table(is.na(res))

# to find out if in clue 2 as well
clue2id<-clue1$CLUE2[t4]
table(is.na(clue2id))

#Check to make sure what these data look like using stem plots, and
make sure they are continuous, binary, or categorical: (Although only
one is shown, this was done for all covariates.
```

```
stem(dtosite)

# Make dataframe with all outcome variables and covariates

c1data2<-data.frame(clue1id,clue2id,addX,addY,dde,ddelip,tpcbspeak,
tpcbspeaklip,achlordane,achlordanelip,gchlordane,gchlordanelip,oxychlor
oxychlorlip,tnonachlor,tnonachlorlip,heptachlor,heptachlorlip,heptepox,
heptepoxlip,HCB,HCBlip,mirex,mirexlip,aldrin,aldrinlip,dieldrin,
dieldrinlip,endrin,endrinlip,bHCH,bHCHlip,gHCH,gHCHlip,dtosite,age,edu,
marital,BMI,race,gender,water,veg,meat,eggs,dairy,fish,occ,alcohol,SES,
zip,smoke2,smoke3,smoke,brfeed,pestpets,pestfarm,pestout,pestin,
pestpest,res,ddt,ddtlip,ddd,dddlip)

# Need to make categorical variables categorical
c1data2$water<-as.factor(c1data2$water)
c1data2$marital<-as.factor(c1data2$marital)
c1data2$fish<-as.factor(c1data2$fish)
c1data2$smoke2<-as.factor(c1data2$smoke2)
c1data2$smoke3<-as.factor(c1data2$smoke3)

# Check data to make sure data matched okay

c1data2[c(101,82,215),]
oc1[c(15,400,1290),]
c1data2[,c(1,17,21,26)]

# Now make some additional variables for possible analysis
# First is a few dtosite buffer variables

# use this as 5 mile buffers away from site
# make buffer in 5 mile radius from site
# (1=<5mi,2=5-10mi,3=10-15mi,4=15-20mi,5=20-25mi,6=25-30mi,7=>30mi)
x<-c1data2$dtosite*5280
y<-c1data2$dtosite*5280
x[y<=26400]<-1
x[y>26400&y<=52800]<-2
x[y>52800&y<=79200]<-3
x[y>79200&y<=105600]<-4
x[y>105600&y<=132000]<-5
x[y>132000&y<=158400]<-6
x[y>158400]<-7
c1data2<-data.frame(c1data2,dbuff2=x)

# and Make rural vs. urban variable
# get point of city origin
HagX<-1109576.10384
HagY<-720505.27649

# Make continuous variable as distance to city (dtocity) in miles

dtocity<-(sqrt((c1data2$addX-HagX)^2+(c1data2$addY-HagY)^2))/5280
c1data2<-data.frame(c1data2,dtocity=dtocity)

#Make Urban vs. Rural variable where rural is anything greater than
8000 ft (1.5miles) from
# center of city
```

```
x<-c1data2$dtocity
y<-c1data2$dtocity
x[y<=1.5]<-1
x[y>1.5]<-2

c1data2<-data.frame(c1data2,urban=x)
stem(c1data2$dbuff2)
stem(c1data2$dtocity)
stem(c1data2$urban)

sitetocenter<-(sqrt((c1data2$sitex-HagX)^2+(c1data2$sitey-
HagY)^2))/5280

# Make a direction variable to see if direction around site makes a
difference, and rotate every 15 degrees
```

Using DirBins function defined as:

```
DirBins<-function(data,rot=0){
  dsite<-data$dtosite*5280
  x<-data$addX
  y<-data$addY
  dx<-x-sitex
  dy<-y-sitey
  n<-length(x)
  quad<-numeric(n)
  quad[dx>0&dy>0]<-1
  quad[dx>0&dy<0]<-2
  quad[dx<0&dy<0]<-3
  quad[dx<0&dy>0]<-4
  angle<-rep(NA,n)
  angle[quad==1]<-asin(abs(dx[quad==1])/dsite[quad==1])
  angle[quad==2]<-90*pi/180 + asin(abs(dy[quad==2])/dsite[quad==2])
  angle[quad==3]<-pi + asin(abs(dx[quad==3])/dsite[quad==3])
  angle[quad==4]<-270*pi/180 + asin(abs(dy[quad==4])/dsite[quad==4])
  angle<-angle*180/pi
  angle<-angle+rot
  bin<-rep(NA,n)
  bin[angle>0 & angle <= 45]<-1
  bin[angle>45 & angle <= 90]<-2
  bin[angle>90 & angle <= 135]<-3
  bin[angle>135 & angle <= 180]<-4
  bin[angle>180 & angle <= 225]<-5
  bin[angle>225 & angle <= 270]<-6
  bin[angle>270 & angle <= 315]<-7
  bin[angle>315 & angle <= 360]<-8
  return(bin)
}

dbins0<-DirBins(c1data2)
c1data2<-data.frame(c1data2,dbins0=dbins0)
cor(c1data2$dde[c1data2$dbins0==1],c1data2$dtosite[c1data2$dbins0==1],u
se="pairwise.complete.obs")
for(i in 1:8)
print(cor(c1data2$dde[c1data2$dbins0==i],c1data2$dtosite[c1data2$dbins0
==i],use="pairwise.complete.obs"))
```

```
dbins15<-DirBins(c1data2,rot=15)
c1data2<-data.frame(c1data2,dbins15=dbins15)
for(i in 1:8)
print(cor(c1data2$dde[c1data2$dbins15==i],c1data2$dtosite[c1data2$dbins
15==i],use="pairwise.complete.obs"))

dbins30<-DirBins(c1data2,rot=30)
c1data2<-data.frame(c1data2,dbins30=dbins30)
for(i in 1:8)
print(cor(c1data2$dde[c1data2$dbins30==i],c1data2$dtosite[c1data2$dbins
30==i],use="pairwise.complete.obs"))

# Make those that need to be categorical, categorical
c1data2$dbuff2<-as.factor(c1data2$dbuff2)
c1data2$dbins0<-as.factor(c1data2$dbins0)
c1data2$dbins15<-as.factor(c1data2$dbins15)
c1data2$dbins30<-as.factor(c1data2$dbins30)

# First made sure no multiple subjects at the same location.  Did this
by adding 1 ft to both the x and y component of their residential
location, then marking a 0,1,2,3,etc in the "changeft" column to
signify if I added 0,1,2,3...ft to each the x and y location,
respectively

# Now there are 4 people in clue 1 that participated in 2 studies. Same
person, same blood, same location. I will average their outcome levels.

#1

c1data2[c1data2$clue1id=="5047",]
c1data2$clue1id[288]<-"5047a"
c1data2$clue1id[289]<-"5047b"
c1data2[288:289,]
c1data2$clue1id[c1data2$clue1id=="5047b"]<-"5047"

u<-c1data2$dde[288]
v<-c1data2$dde[289]
c1data2$dde[c1data2$clue1id=="5047"]<-mean(c(u,v))
u<-c1data2$ddelip[288]
v<-c1data2$ddelip[289]
c1data2$ddelip[c1data2$clue1id=="5047"]<-mean(c(u,v))

u<-c1data2$tpcbspeak[288]
v<-c1data2$tpcbspeak[289]
c1data2$tpcbspeak[c1data2$clue1id=="5047"]<-mean(c(u,v))
u<-c1data2$tpcbspeaklip[288]
v<-c1data2$tpcbspeaklip[289]
c1data2$tpcbspeaklip[c1data2$clue1id=="5047"]<-mean(c(u,v))

dim(c1data2)

a<-c(1:1391)[c1data2$clue1id=="5047a"]
c1data2<-c1data2[-a,]
c1data2[c1data2$clue1id=="5047",]

#2
```

```
c(1:1390)[c1data2$clue1id=="31033"]
c1data2[c1data2$clue1id=="31033",]
c1data2$clue1id[907]<-"31033a"
c1data2$clue1id[908]<-"31033b"
c1data2[907:908,]

c1data2$clue1id[c1data2$clue1id=="31033b"]<-"31033"

u<-c1data2$dde[907]
v<-c1data2$dde[908]
c1data2$dde[c1data2$clue1id=="31033"]<-mean(c(u,v))
u<-c1data2$ddelip[907]
v<-c1data2$ddelip[908]
c1data2$ddelip[c1data2$clue1id=="31033"]<-mean(c(u,v))

u<-c1data2$tpcbspeak[907]
v<-c1data2$tpcbspeak[908]
c1data2$tpcbspeak[c1data2$clue1id=="31033"]<-mean(c(u,v))
u<-c1data2$tpcbspeaklip[907]
v<-c1data2$tpcbspeaklip[908]
c1data2$tpcbspeaklip[c1data2$clue1id=="31033"]<-mean(c(u,v))

dim(c1data2)

b<-c(1:1390)[c1data2$clue1id=="31033a"]
c1data2<-c1data2[-b,]
c1data2[c1data2$clue1id=="31033",]

#3
c(1:1389)[c1data2$clue1id=="33872"]
c1data2[c1data2$clue1id=="33872",]
c1data2$clue1id[1056]<-"33872a"
c1data2$clue1id[1057]<-"33872b"
c1data2[1056:1057,]

c1data2$clue1id[c1data2$clue1id=="33872b"]<-"33872"

u<-c1data2$dde[1056]
v<-c1data2$dde[1057]
c1data2$dde[c1data2$clue1id=="33872"]<-mean(c(u,v))
u<-c1data2$ddelip[1056]
v<-c1data2$ddelip[1057]
c1data2$ddelip[c1data2$clue1id=="33872"]<-mean(c(u,v))

u<-c1data2$tpcbspeak[1056]
v<-c1data2$tpcbspeak[1057]
c1data2$tpcbspeak[c1data2$clue1id=="33872"]<-mean(c(u,v))
u<-c1data2$tpcbspeaklip[1056]
v<-c1data2$tpcbspeaklip[1057]
c1data2$tpcbspeaklip[c1data2$clue1id=="33872"]<-mean(c(u,v))

c<-c(1:1389)[c1data2$clue1id=="33872a"]
c1data2<-c1data2[-c,]
c1data2[c1data2$clue1id=="33872",]

#4
c(1:1388)[c1data2$clue1id=="34913"]
```

```
cldata2[cldata2$cluelid=="34913",]
cldata2$cluelid[1110]<-"34913a"
cldata2$cluelid[1111]<-"34913b"
cldata2[1110:1111,]

cldata2$cluelid[cldata2$cluelid=="34913b"]<-"34913"

u<-cldata2$dde[1110]
v<-cldata2$dde[1111]
cldata2$dde[cldata2$cluelid=="34913"]<-mean(c(u,v))
u<-cldata2$ddelip[1110]
v<-cldata2$ddelip[1111]
cldata2$ddelip[cldata2$cluelid=="34913"]<-mean(c(u,v))

u<-cldata2$tpcbspeak[1110]
v<-cldata2$tpcbspeak[1111]
cldata2$tpcbspeak[cldata2$cluelid=="34913"]<-mean(c(u,v))
u<-cldata2$tpcbspeaklip[1110]
v<-cldata2$tpcbspeaklip[1111]
cldata2$tpcbspeaklip[cldata2$cluelid=="34913"]<-mean(c(u,v))

dim(cldata2)

d<-c(1:1388)[cldata2$cluelid=="34913a"]
cldata2<-cldata2[-d,]
cldata2[cldata2$cluelid=="34913",]

# Take out variables that aren't in the WC zip codes
u<-c(1:1387)[cldata2$zip==21736|cldata2$zip==25419|cldata2$zip==25425|
cldata2$zip==21712]
cldata2<-cldata2[-u,]

# Take out 4 points where geocoded location is not inside washington
County
dim(cldata2)
v<-c(1:1379)[cldata2$cluelid==41266|cldata2$cluelid==12637|
cldata2$cluelid==13849|cldata2$cluelid==13854]
cldata2<-cldata2[-v,]

# try w/o outlier

names(cldata2)
dim(cldata2)
u<-c(1:1375)[cldata2$dde>400]
cldata2<-data.frame(cldata2[-u,])
dim(cldata2)

# Recode Water to only 4 categories: 0=city water, 1=spring water,
2=well water, 3=other, Recode to make city water the reference
category: 0=city,1=spring,2=well,3=cistern (rain, mixed, or
other),4=bottled

n<-length(cldata2$water)
water2<-rep(NA,n)
cldata2<-data.frame(cldata2,water2=water2)
cldata2$water2[cldata2$water==0]<-0
cldata2$water2[cldata2$water==1]<-1
```

```
c1data2$water2[c1data2$water==2]<-2
c1data2$water2[c1data2$water==3|c1data2$water==4]<-3

water2<-c1data2$water2
table(water2)
c1data2$water2<-as.factor(c1data2$water2)

# try recoding gender here to match clue 2

# recode to match clue 2 coding 1=female, 2=male

n<-length(c1data2$gender)
gender2<-rep(NA,n)
c1data2<-data.frame(c1data2,gender2=gender2)
c1data2$gender2[c1data2$gender==2]<-1
c1data2$gender2[c1data2$gender==1]<-2

gender2<-c1data2$gender2
table(c1data2$gender2)
table(c1data2$gender)

# Descriptive statistics on outcome variables (w/o zeros)

Mdde<-mean.default(c1data2$dde,na.rm=T)
Ndde<-length(c1data2$dde[!is.na(c1data2$dde)])
SDdde<-sqrt(var(c1data2$dde,na.rm=T))/sqrt(Ndde)
CILdde<-Mdde-1.96*SDdde
CIUdde<-Mdde+1.96*SDdde
Mdde
Ndde
SDdde
CILdde
CIUdde
summary(c1data2$dde)

# Make new data frame ready to make geodata object with

library("geoR")

geodata<-data.frame(addX=c1data2$addX,addY=c1data2$addY,dde=
c1data2$dde,ddelip=c1data2$ddelip,tpcbspeak=c1data2$tpcbspeak,tpcbspeak
lip= c1data2$tpcbspeaklip,clue1id=c1data2$clue1id,achlordane=
c1data2$achlordane,achlordanelip=c1data2$achlordanelip,
gchlordane=c1data2$gchlordane,gchlordanelip=c1data2$gchlordanelip,
oxychlor=c1data2$oxychlor,oxychlorlip=c1data2$oxychlorlip,tnonachlor=
c1data2$tnonachlor,tnonachlorlip=c1data2$tnonachlorlip,heptachlor=
c1data2$heptachlor,heptachlorlip=c1data2$heptachlorlip,heptepox=
c1data2$heptepox,heptepoxlip=c1data2$heptepoxlip,HCB=c1data2$HCB,
HCBlip=c1data2$HCBlip,mirex=c1data2$mirex,mirexlip=c1data2$mirexlip,
aldrin=c1data2$aldrin,aldrinlip=c1data2$aldrinlip,dieldrin=
c1data2$dieldrin,dieldrinlip=c1data2$dieldrinlip,endrin=c1data2$endrin,
endrinlip=c1data2$endrinlip,bHCH=c1data2$bHCH,bHCHlip=c1data2$bHCHlip,
gHCH=c1data2$gHCH,gHCHlip=c1data2$gHCHlip,ddt=c1data2$ddt,ddtlip=
c1data2$ddtlip,ddd=c1data2$ddd,dddlip=c1data2$dddlip)

# First, clean up data (NAs in coordinates or outcome variable)
# And make them into geodata objects forVariogram analyses for OC data
# Taking out missing data for DDE
dim(geodata)
u1<-c(1:1374)[is.na(geodata$dde)]
u2<-c(1:1374)[geodata$dde<=0]
u3<-c(1:1374)[is.na(geodata$addX)|is.na(geodata$addY)]
u<-c(u1,u2,u3)
u<-unique(u)

# Make geodata objects without the missing data

ddegeo<-as.geodata(geodata[-u,],data.col=3,covar.col=7)

# Write these into text files to bring them into GIS: R->.csv->GIS

write.table(geodata[-u,],file="c:\\shenshaw\\Thesis\\CLUE\\CLUE_Data\\
Changed_Variable_Names\\Merged_Data\\ddegeodata.csv",sep=",",col.names=
T,row.names=F)

#Exaustive regression search looking for most influential risk factors
```

This is an example of the exhaustive search of one organochlorine, DDE. However, this was done for each organochlorine, both adjusted and unadjusted for lipids.

Furthermore, only one loop is shown, and all combinations run (but not shown are described at the end of the loop.

```
library("combinat")
# Make new dataset with only those variables I wish to test together

dde75a<-
c1data[,c("dde","dtosite","age","edu","marital","race","gender","water"
,"smoke2","SES","dbins0")]

# Have to run this code each time and change data name

RunReg<-function(vars,resp="dde"){
  xnam<-names(dde75a)[vars]
  fmla<-as.formula(paste(resp," ~ ", paste(xnam,collapse=" + ")))
  fit<-lm(fmla,data=dde75a)
  rsqrd<-round(100*summary(fit)$adj.r.squared,2)
  fmla<-as.character(fmla)
  fmla<-paste(fmla[2],paste(" ",fmla[1]," ",sep=""),fmla[3],sep="")
  return(c(rsqrd,as.character(fmla)))
}

# This has 11 possible variables and one response listed in the first
column. Below you'll need t1 through tN, where N is the number of
possible variables

models<-combn(2:11,1)
t1<-apply(models,2,RunReg)

models<-combn(2:11,2)
t2<-apply(models,2,RunReg)
```

```
models<-combn(2:11,3)
t3<-apply(models,2,RunReg)

models<-combn(2:11,4)
t4<-apply(models,2,RunReg)

models<-combn(2:11,5)
t5<-apply(models,2,RunReg)

models<-combn(2:11,6)
t6<-apply(models,2,RunReg)

models<-combn(2:11,7)
t7<-apply(models,2,RunReg)

models<-combn(2:11,8)
t8<-apply(models,2,RunReg)

models<-combn(2:11,9)
t8a<-apply(models,2,RunReg)

models<-matrix(combn(2:11,10),ncol=1)
t9<-apply(models,2,RunReg)

tall<-cbind(t1,t2,t3,t4,t5,t6,t7,t8,t8a,t9)
Results<-data.frame(t(tall))
names(Results)<-c("AdjRsqrd","Model")
Results$AdjRsqrd<-as.numeric(as.vector(Results$AdjRsqrd))
u<-order(Results$AdjRsqrd,decreasing=TRUE)
Results[u[1:10],]

# b Try using addX,addY instead of dtosite

dde75b<-
c1data[,c("dde","addX","addY","age","edu","marital","race","gender","wa
ter","smoke2","SES","dbins0")]

#c Try using distance buffer instead of distance to site/addX/addY

dde75c<-
c1data[,c("dde","dbuff2","age","edu","marital","race","gender","water",
"smoke2","SES","dbins0")]

#d Try using urban/rural variable instead of other distance to site
variables

dde75d<-
c1data[,c("dde","urban","age","edu","marital","race","gender","water","
smoke2","SES","dbins0")]

# Try doing steps, a, b, c and d with smoke as binary, see if makes a
difference:

# e Make new dataset with only those variables I wish to test together
```

```
dde75e<-
c1data[,c("dde","dtosite","age","edu","marital","race","gender","water"
,"smoke","SES","dbins0")]

# f Try using addX,addY instead of dtosite

dde75f<-
c1data[,c("dde","addX","addY","age","edu","marital","race","gender","wa
ter","smoke","SES","dbins0")]

#g Try using distance buffer instead of distance to site/addX/addY

dde75g<-
c1data[,c("dde","dbuff2","age","edu","marital","race","gender","water",
"smoke","SES","dbins0")]

#h Try using urban/rural variable instead of other distance to site
variables

dde75h<-
c1data[,c("dde","urban","age","edu","marital","race","gender","water","
smoke","SES","dbins0")]

#Now do steps a-h with direction bins rotated by 15 degrees

dde75a15<-
c1data[,c("dde","dtosite","age","edu","marital","race","gender","water"
,"smoke2","SES","dbins15")]

#Now do steps a-h with direction bins rotated by 30 degrees

dde75a30<-
c1data[,c("dde","dtosite","age","edu","marital","race","gender","water"
,"smoke2","SES","dbins30")]

# Now combine all models get results of all of these combinations
ordered according to R2:

tall<-
cbind(t1,t2,t3,t4,t5,t6,t7,t8,t8a,t9,t10,t11,t12,t13,t14,t15,t16,t17,t1
8,t18a,t19,t20,t21,t22,t23,t24,t25,t26,t27,t27a,t28,t29,t30,t31,t32,
t33,t34,t35,t36,t36a,t37,t38,t39,t40,t41,t42,t43,t44,t45,t45a,t46,t47,
t48,t49,t50,t51,t52,t53,t54,t55,t55a,t56,t57,t58,t59,t60,t61,t62,t63,
t64,t64a,t65,t66,t67,t68,t69,t70,t71,t72,t73,t73a,t74,t101,t102,t103,
t104,t105,t106,t107,t108,t108a,t109,t110,t111,t112,t113,t114,t115,t116,
t117,t118,t118a,t119,t120,t121,t122,t123,t124,t125,t126,t127,t127a,
t128,t129,t130,t131,t132,t133,t134,t135,t136,t136a,t137,t138,t139,t140,
t141,t142,t143,t144,t145,t145a,t146,t147,t148,t149,t150,t151,t152,t153,
t154,t155,t155a,t156,t157,t158,t159,t160,t161,t162,t163,t164,t164a,
t165,t166,t167,t168,t169,t170,t171,t172,t173,t173a,t174,t201,t202,t203,
t204,t205,t206,t207,t208,t208a,t209,t210,t211,t212,t213,t214,t215,t216,
t217,t218,t218a,t219,t220,t221,t222,t223,t224,t225,t226,t227,t227a,
t228,t229,t230,t231,t232,t233,t234,t235,t236,t236a,t237,t238,t239,t240,
t241,t242,t243,t244,t245,t245a,t246,t247,t248,t249,t250,t251,t252,t253,
t254,t255,t255a,t256,t257,t258,t259,t260,t261,t262,t263,t264,t264a,
t265,t266, t267,t268,t269,t270,t271,t272,t273,t273a,t274)
Results<-data.frame(t(tall))
```

```
names(Results)<-c("AdjRsqrd","Model")
Results$AdjRsqrd<-as.numeric(as.vector(Results$AdjRsqrd))
u<-order(Results$AdjRsqrd,decreasing=TRUE)
Results[u[1:40],]

write(t(Results[u,]),file="c:\\shenshaw\\Thesis\\Papers\\Paper3\\dde74.
txt",ncol=2)

# Best OLS models chosen for each organochlorine:

# use only whites
c1data3<-c1data2[c1data2$race==1,]

ddereg<-lm(dde ~ dtosite + age +  gender2 + water2 + smoke + edu,
c1data3)
summary(ddereg)

#####################################################################

# all without spatial

ddenospatial<-lm(dde ~ age +  gender2 + water2 + smoke + edu
      +race*edu, c1data3)
summary(ddenospatial)

##################################################################

# Analyzing models
# DDE

plot(variog(ddegeo,max.dist=100000))
title(" variogram for DDE ")
plot(variog4(ddegeo,max.dist=100000))
title("multidirectional variogram for DDE ")

# logged

plot(variog(ddegeo,max.dist=100000,lambda=0))
title(" variogram for logged DDE ")
plot(variog4(ddegeo,max.dist=100000,lambda=0))
title("multidirectional variogram for logged DDE ")

#####################################################################

# plot variograms of residuals of models:
# dde

u<-complete.cases(c1data3[,c("dde","addX","addY","dtosite",
"age","race","gender2","water2","smoke","SES")])
ddereggeo<-as.geodata(cbind(c1data3[u,c("addX","addY")],
as.vector(ddereg$residuals)))
plot(variog(ddereggeo,max.dist=100000))
title("V residuals of dde 74 model")
plot(variog4(ddereggeo,max.dist=100000))
title("Multidirectional V residuals of dde 74 model")

# check to make sure lined up alright
```

```
t1<-c1data3[u,c("addX","addY")]
t2<-ddereg$residuals
range(as.numeric(labels(t1)[[1]])-as.numeric(labels(t2)))

# dde no spatial

ddenospatialb<-lm(dde ~ age + race + gender2 + water2 + smoke +SES +
race*SES,c1data3[u,])
ddenospatialgeo<-as.geodata(cbind(c1data3[u,c("addX","addY")],
as.vector(ddenospatialb$residuals)))
plot(variog(ddenospatialgeo,max.dist=100000))
title("V residuals of dde 74 model w/o spatial")
plot(variog4(ddenospatialgeo,max.dist=100000))
title("Multidirectional V residuals of dde 74 model w/o spatial")

##############fitting variogram models ##################

# now fit exponential model to residual variograms

# DDE

plot(variog(ddereggeo))
plot(variog(ddereggeo,max.dist=100000))
lines.variomodel(cov.model="gaussian",cov.pars=
c(150,30000),nugget=150,max.dist=100000)
title("Estimating a Variogram model for 74 dde  data")
dde.remla<-likfit(ddereggeo,ini=c(150,30000),nugget=150,cov.model =
"gaussian",method="reml")
names(dde.remla)
dde.remla[1:3]

# look at all variables to see what is going on, also done with race in
a separate analysis, but not shown here since not reported in
Manuscript Two

# outcome variables

hist(c1data3$dde)
hist(c1data3$dtosite)

# Look at Covariates using the table and summary commands (only example
is shown)

table(c1data3$gender2)
summary(c1data3$dtosite)

# check correlations,

cor(c1data3$SES,c1data3$edu, use="pairwise.complete.obs")
cor(c1data3[,39:72], use="pairwise.complete.obs")

# Look at data stratified by all covariats (only gender shown)

table(c1data3$gender2,c1data3$water2)
table(c1data3$gender2,c1data3$marital)
table(c1data3$gender2,c1data3$smoke)
```

```
# make new summary statistics stratifying by gender, (also done with
race, urban/rural residence and smoking status, but not shown):

# gender

by(c1data3$dde, c1data3$gender2, summary, na.rm=T)
by(c1data3$ddelip, c1data3$gender2, summary, na.rm=T)
by(c1data3$dtosite, c1data3$gender2, summary, na.rm=T)
by(c1data3$SES, c1data3$gender2, summary, na.rm=T)
by(c1data3$edu, c1data3$gender2, summary, na.rm=T)
by(c1data3$age, c1data3$gender2, summary, na.rm=T)
by(c1data3$urban, c1data3$gender2, summary, na.rm=T)
by(c1data3$smoke, c1data3$gender2, summary, na.rm=T)
by(c1data3$gender2, c1data3$gender2, summary, na.rm=T)
by(c1data3, c1data3$gender2, mean, na.rm=T)

# Now generalized Least square regression was used taking into account
the spatial dependence in the data was fit to each organochlorine
chosen to represent CLUE I analysis.

# Producing GLS regression correction results CLUE 1

######### DDE ##################################

u<-complete.cases(c1data3[,c("dde","addX","addY","dtosite","age",
"gender2","water2","smoke","edu")])
tempdde<-c1data3[u,]
names(tempdde)
covarsdde<-model.matrix(~dtosite+age+gender2+edu+water2+smoke,
data=tempdde)
covarsdde<-data.frame(covarsdde)
covarsdde<-covarsdde[,-1]
# 5th column is dde, x, y are 3,4
temp2dde<-data.frame(tempdde[,c(3:4,5)],covarsdde)
t2geodde<-as.geodata(temp2dde,coords.col=1:2,data.col=3,covar.col=4:11)

vg2adde<-variog(t2geodde,max.dist=100000)
plot(vg2adde)

trend1dde<-formula(~t2geodde$covariate$dtosite+t2geodde$covariate$age+
t2geodde$covariate$gender2+t2geodde$covariate$edu+
t2geodde$covariate$water21+t2geodde$covariate$water22+
t2geodde$covariate$water23+t2geodde$covariate$smoke)

vg2bdde<-variog(t2geodde,max.dist=100000,trend=trend1dde)
plot(vg2bdde)
vg2bdde$beta.ols

# check to make sure these match the betas for lm method

ddelmtemp2<-lm(dde~dtosite+age+gender2+edu+water2+smoke,data=tempdde)
ddelmtemp2$coefficients

locs1dde<-t2geodde$coords

# initial parameters were estimated based on the residual variation
found from the OLS regression
```

```
cov1<-c(200,30000)
nugget1<-150
n<-dim(tempdde)[1]

fit2bdde<-likfit(t2geodde, cov.model="exp", ini.cov.pars=cov1,
method="reml",nugget=nugget1, trend=trend1dde)

plot(vg2bdde)
lines.variomodel(fit2bdde)

# beta GLS coef. and GLS std error
glsddea<-cbind(fit2bdde$beta,sqrt(diag(fit2bdde$beta.var)),
fit2bdde[[7]]/(sqrt(diag(fit2bdde[[8]]))))
glsddea

# get R2 and adjusted R2

dmatdde<-as.matrix(dist(temp2dde[,1:2]))
Cov<-cov.spatial(dmatdde,cov.model="exp",cov.pars=fit2bdde$cov.pars)
Cov<-Cov/fit2bdde$cov.pars[1] # so Now we have (sigma^2)V as covariance
structure
Cinv<-solve(Cov)
X<-as.matrix(cbind(1,temp2dde[,4:11]))
p<-dim(X)[2]
Y<-matrix(temp2dde[,3],ncol=1)
beta.ols<-solve(t(X)%*%X)%*%t(X)%*%Y
var.ols<-diag(solve(t(X)%*%X))
sse.ols<-t((Y-X%*%beta.ols))%*%(Y-X%*%beta.ols)
mse.ols<-sse.ols/(n-p)
ols<-cbind(beta.ols,sqrt(mse.ols*var.ols))
sqrt(mse.ols)

# matches residual standard error from summary(temp2)
summary(temp2dde)

beta.gls<-solve(t(X)%*%Cinv%*%X)%*%t(X)%*%Cinv%*%Y
var.gls<-diag(solve(t(X)%*%Cinv%*%X))
sse.gls<-t((Y-X%*%beta.gls))%*%Cinv%*%(Y-X%*%beta.gls)
mse.gls<-sse.gls/(n-p)
gls<-cbind(beta.gls,sqrt(mse.gls*var.gls))

J<-matrix(1,n,1)
ssto.gls<-t(Y)%*%Cinv%*%Y-
solve(t(J)%*%Cinv%*%J)%*%(t(Y)%*%Cinv%*%J%*%t(J)%*%Cinv%*%Y)
ssr.gls<-ssto.gls-sse.gls
R2.gls<-1-sse.gls/ssto.gls
R2adj.gls<-1-((n-1)/(n-p))*sse.gls/ssto.gls

e.ols<-Y-X%*%beta.ols
e.gls<-Y-X%*%beta.gls

muhat<-X%*%beta.gls
resids<-Y-muhat

ddegeo2<-as.geodata(cbind(t2geodde$coords,resids))
```

```
plot(variog(ddegeo2))
plot(variog(ddegeo2, max.dist=100000))

# Now re-estimate initial parameters from residual variation in GLS
regressionas (a 2nd iteration).

cov1<-c(220,30000)
nugget1<-180

fit2bdde2<-likfit(t2geodde, cov.model="exp", ini.cov.pars=cov1,
method="reml",nugget=nugget1, trend=trend1dde)

glsddeb<-
cbind(fit2bdde2$beta,sqrt(diag(fit2bdde2$beta.var)),fit2bdde2[[7]]/(sqr
t(diag(fit2bdde2[[8]]))))
glsddeb

# Get R2 and adjusted R2 for this final model

dmatdde<-as.matrix(dist(temp2dde[,1:2]))
Cov<-cov.spatial(dmatdde,cov.model="exp",cov.pars=fit2bdde2$cov.pars)
Cov<-Cov/fit2bdde2$cov.pars[1] # so Now we have (sigma^2)V as
covariance structure
Cinv<-solve(Cov)
X<-as.matrix(cbind(1,temp2dde[,4:11]))
p<-dim(X)[2]
Y<-matrix(temp2dde[,3],ncol=1)
beta.ols<-solve(t(X)%*%X)%*%t(X)%*%Y
var.ols<-diag(solve(t(X)%*%X))
sse.ols<-t((Y-X%*%beta.ols))%*%(Y-X%*%beta.ols)
mse.ols<-sse.ols/(n-p)
ols<-cbind(beta.ols,sqrt(mse.ols*var.ols))
sqrt(mse.ols)

# matches residual standard error from summary(temp2)
summary(temp2dde)

beta.gls<-solve(t(X)%*%Cinv%*%X)%*%t(X)%*%Cinv%*%Y
var.gls<-diag(solve(t(X)%*%Cinv%*%X))
sse.gls<-t((Y-X%*%beta.gls))%*%Cinv%*%(Y-X%*%beta.gls)
mse.gls<-sse.gls/(n-p)
gls<-cbind(beta.gls,sqrt(mse.gls*var.gls))

J<-matrix(1,n,1)
ssto.gls<-t(Y)%*%Cinv%*%Y-
solve(t(J)%*%Cinv%*%J)%*%(t(Y)%*%Cinv%*%J%*%t(J)%*%Cinv%*%Y)
ssr.gls<-ssto.gls-sse.gls
R2.gls2<-1-sse.gls/ssto.gls
R2adj.gls2<-1-((n-1)/(n-p))*sse.gls/ssto.gls

glsdderesults1<-cbind(R2.gls,R2adj.gls,R2.gls2,R2adj.gls2)
glsdderesults1
glsdderesults<-cbind(glsddea, glsddeb)

# Run simple, univariate (crude) analyses using GLS regression
```

```
fit2bdde1<-likfit(t2geodde, cov.model="exp", ini.cov.pars=cov1,
method="reml",nugget=nugget1, trend=~t2geodde$covariate$dtosite)
glsdde1<-c(fit2bdde1$beta,sqrt(diag(fit2bdde1$beta.var)),
fit2bdde1$beta/(sqrt(diag(fit2bdde1$beta.var))))

fit2bdde2<-likfit(t2geodde, cov.model="exp", ini.cov.pars=cov1,
method="reml",nugget=nugget1, trend=~t2geodde$covariate$age)
glsdde2<-c(fit2bdde2$beta,sqrt(diag(fit2bdde2$beta.var)),
fit2bdde2$beta/(sqrt(diag(fit2bdde2$beta.var))))

fit2bdde3<-likfit(t2geodde, cov.model="exp", ini.cov.pars=cov1,
method="reml",nugget=nugget1, trend=~t2geodde$covariate$gender2)
glsdde3<-c(fit2bdde3$beta,sqrt(diag(fit2bdde3$beta.var)),
fit2bdde3$beta/(sqrt(diag(fit2bdde3$beta.var))))

fit2bdde5<-likfit(t2geodde, cov.model="exp", ini.cov.pars=cov1,
method="reml",nugget=nugget1, trend=~t2geodde$covariate$edu)
glsdde5<-c(fit2bdde5$beta,sqrt(diag(fit2bdde5$beta.var)),
fit2bdde5$beta/(sqrt(diag(fit2bdde5$beta.var))))

fit2bdde6<-likfit(t2geodde, cov.model="exp", ini.cov.pars=cov1,
method="reml",nugget=nugget1, trend=~t2geodde$covariate$smoke)
glsdde6<-c(fit2bdde6$beta,sqrt(diag(fit2bdde6$beta.var)),
fit2bdde6$beta/(sqrt(diag(fit2bdde6$beta.var))))

fit2bdde7<-likfit(t2geodde, cov.model="exp", ini.cov.pars=cov1,
method="reml",nugget=nugget1, trend=~t2geodde$covariate$water21)
glsdde7<-c(fit2bdde7$beta,sqrt(diag(fit2bdde7$beta.var)),
fit2bdde7$beta/(sqrt(diag(fit2bdde7$beta.var))))

fit2bdde8<-likfit(t2geodde, cov.model="exp", ini.cov.pars=cov1,
method="reml",nugget=nugget1, trend=~t2geodde$covariate$water22)
glsdde8<-c(fit2bdde8$beta,sqrt(diag(fit2bdde8$beta.var)),
fit2bdde8$beta/(sqrt(diag(fit2bdde8$beta.var))))

fit2bdde9<-likfit(t2geodde, cov.model="exp", ini.cov.pars=cov1,
method="reml",nugget=nugget1, trend=~t2geodde$covariate$water23)
glsdde9<-c(fit2bdde9$beta,sqrt(diag(fit2bdde9$beta.var)),
fit2bdde9$beta/(sqrt(diag(fit2bdde9$beta.var))))

simresultsdde<-rbind(glsdde1,glsdde2,glsdde3,glsdde5,glsdde6,glsdde7,
                     glsdde8,glsdde9)

write.table(simresultsdde,file="c:\\shenshaw\\Thesis\\Papers\\Paper3\\g
ls\\simdde.csv",sep=",",col.names=T,row.names=F)
write.table(glsdderesults,file="c:\\shenshaw\\Thesis\\Papers\\Paper3\\g
ls\\glsdde.csv",sep=",",col.names=T,row.names=F)
write.table(glsdderesults1,file="c:\\shenshaw\\Thesis\\Papers\\Paper3\\
gls\\glsdde1.csv",sep=",",col.names=T,row.names=F)
```

CLUE II (1989) Analysis

Like CLUE I (1974), only the analysis of one compound, DDE levels in plasma, are shown here. All other organochlorines both lipid adjusted and unadjusted were also analyzed in the same manner.

```
###  Reading in all of the data ###

#  Organochlorine Data

# Clue 2: 1989
oc2<-read.csv("c:\\shenshaw\\Thesis\\CLUE\\CLUE_Data\\
Changed_Variable_Names\\Merged_Data\\wcc2ocdNa.csv",header=T,sep=",")

# Clue 2: prostate cancer lipid data
prlipc2<-read.csv("c:\\shenshaw\\Thesis\\CLUE\\CLUE_Data\\ExtraVars\\
prlipc2.csv",header=T,sep=",")

# Address Data:
# clue 2: 1989
addr2<-read.csv("c:\\shenshaw\\Thesis\\GIS\\Geocode\\clue2\\
clue2shpNAwoxy.dup.csv",header=T,sep=",")

# Extra Variable data

# Breast Cancer Studies 1995
bc1<-read.csv("c:\\shenshaw\\Thesis\\CLUE\\CLUE_Data\\ExtraVars\\
bcsvars1NA.csv",header=T,sep=",")
bc2<-read.csv("c:\\shenshaw\\Thesis\\CLUE\\CLUE_Data\\ExtraVars\\
bcsvars2NA.csv",header=T,sep=",")
bc3<-read.csv("c:\\shenshaw\\Thesis\\CLUE\\CLUE_Data\\ExtraVars\\
bcsvars3NA.csv",header=T,sep=",")
bc4<-read.csv("c:\\shenshaw\\Thesis\\CLUE\\CLUE_Data\\ExtraVars\\
bcsvars4NA.csv",header=T,sep=",")

# CLUE Studies
# Clue 2: 1989
clue2<-read.csv("c:\\shenshaw\\Thesis\\CLUE\\CLUE_Data\\ExtraVars\\
c2vars1NA.csv",header=T,sep=",")

# CLUE Follow up studies 1996
fol96<-read.csv("c:\\shenshaw\\Thesis\\CLUE\\CLUE_Data\\ExtraVars\\
c2med96NA.csv",header=T,sep=",")

# Clue 2 Round 2 Study 1998
rnd2<-read.csv("c:\\shenshaw\\Thesis\\CLUE\\CLUE_Data\\ExtraVars\\
c2rnd2NA.csv",header=T,sep=",")

# 2000
fol00<-read.csv("c:\\shenshaw\\Thesis\\CLUE\\CLUE_Data\\ExtraVars\\
c2med00NA.csv",header=T,sep=",")

# Diet Study: 1989
diet<-read.csv("c:\\shenshaw\\Thesis\\CLUE\\CLUE_Data\\ExtraVars\\
c2diet89NA.csv",header=T,sep=",")
```

```
fruit<-read.csv("c:\\shenshaw\\Thesis\\CLUE\\CLUE_Data\\ExtraVars\\
fruitNA.csv",header=T,sep=",")

# Additional variables asked for from several studies
mixed<-read.csv("c:\\shenshaw\\Thesis\\CLUE\\CLUE_Data\\ExtraVars\\
mixedNA.csv",header=T,sep=",")

### CLUE II only ###

clue2ID<-oc2$clue2
clue1ID<-oc2$clue1

# DDE
table(is.na(oc2$pdde.ml2))
table(is.na(oc2$pdde.lp2))

cor(oc2$pdde.ml2,oc2$oddt.ml2, use="pairwise.complete.obs")
cor(oc2$pdde.ml2,oc2$pddt.ml2, use="pairwise.complete.obs")
cor(oc2$pdde.ml2,oc2$pddd.ml2, use="pairwise.complete.obs")
cor(oc2$pdde.lp2,oc2$oddt.lp2, use="pairwise.complete.obs")
cor(oc2$pdde.lp2,oc2$pddt.lp2, use="pairwise.complete.obs")
cor(oc2$pdde.lp2,oc2$pddd.lp2, use="pairwise.complete.obs")

# Make total lipids variables using cholesterol and triglycerides for
Philips et al short equation..equation 2 then make it a factor to
multiply unadjusted values by to get ng/g units

tt<-match(oc2$clue2,prlipc2$clue2)

choles<-prlipc2$choles[tt]
oc2<-data.frame(oc2,cholesx=choles)

n<-length(oc2$tl1x)
tlp2<-rep(NA,n)
oc2<-data.frame(oc2,tlp2=tlp2)
oc2$tlp2[oc2$study==1]<-oc2$pdde.lp2[oc2$study==1]/
oc2$pdde.ml2[oc2$study==1]
oc2$tlp2[oc2$study==2]<-1/(2.27*oc2$cholesx[oc2$study==2]+
oc2$triglycx[oc2$study==2]+0.623)*100000
oc2$tlp2[oc2$study==3]<-1/(2.27*oc2$cholstrx[oc2$study==3]+
oc2$triglycx[oc2$study==3]+0.623)*100000
tlp2<-oc2$tlp2

# Get a total lipids variable for breast cancer study since cholesterol
and triglyceride data were not available for 1989 breast cancer data.

oc2$pdde.lp2[oc2$study==1]/oc2$pdde.ml2[oc2$study==1]
n<-length(oc2$p153.ml2)
c2bclip2<-rep(NA,n)
oc2<-data.frame(oc2,c2bclip2=c2bclip2)
oc2$c2bclip2[oc2$study==1]<-
oc2$p170.lp2[oc2$study==1]/oc2$p170.ml2[oc2$study==1]
oc2$c2bclip2[oc2$study==2]<-NA
oc2$c2bclip2[oc2$study==3]<-NA

oc2$pdde.lp2<-oc2$pdde.ml2*oc2$tlp2
```

```
# Find zeros in all data

zeropdde<-length(oc2$pdde.ml2[oc2$pdde.ml2==0&!is.na(oc2$pdde.ml2)])
/length(oc2$pdde.ml2[!is.na(oc2$pdde.ml2)])*100

# DDT compounds

# DDE
# To find out where the zeros are:
u<-c(1:797)[oc2$pdde.ml2==0]
# To find out what study the zeros are found in
oc2[u,5]
# Make them the LOD/sqrt(2)
oc2$pdde.ml2[oc2$pdde.ml2==0]<-0.61/sqrt(2)

# get zeros out of lipid adjusted for study 1 (BC) as well
#cannot change lipids to zero for DDE because that is where got factor
from, why had to make bclip2 as a backup

oc2$pdde.lp2[oc2$pdde.lp2==0]<-
oc2$pdde.ml2[oc2$pdde.lp2==0]*oc2$c2bclip2[oc2$pdde.lp2==0]
c2dde<-oc2$pdde.ml2
c2ddelip<-oc2$pdde.lp2

### x1,2,3..<-all covariates ###

t10<-match(oc2$clue2,addr2$CLUE.2)
t12<-match(oc2$clue2,fol96$CLUE2)
t13<-match(oc2$clue2,fol00$CLUE2)
t14<-match(oc2$clue2,clue2$CLUE2)
t15<-match(oc2$clue2,diet$CLUE2)
t16<-match(oc2$clue2,rnd2$CLUE2)
t17<-match(oc2$clue2,c1data$clue2id)
t18<-match(oc2$clue2,bc1$Clue)
t19<-match(oc2$clue2,bc2$Clue)
t20<-match(oc2$clue2,bc4$CLUE2)
t21<-match(oc2$clue2,mixed$Clue2)
t22<-match(oc2$clue2,fruit$CLUE2)

### Continuous Variables..###

# Centroid of site in MD state plane coords 83 ft
sitex<-1108776
sitey <-725732

residence: zip and x and y location in MD state plane coordinates
c2zip<-addr2$ARC.ZONE[t10]
c2addX<-addr2$MD83ft.X[t10]
c2addY<-addr2$MD83ft.Y[t10]

# Distance from residence location to site measured in feet
c2dtosite<-(sqrt((c2addX-sitex)^2+(c2addY-sitey)^2))/5280

# Age in years
c2age<-clue2$Age[t14]
```

```
# Body Mass Index
c2BMI<-clue2$BMI[t14]

# Number of years of school completed
c2edu<-clue2$Ed[t14]

# Months of breastfeeding
c2brfeed<-bc1$Breastfd[t18]

# Total years applied insecticides on pets
c2pestpets<-bc2$petyrs[t19]

# Total years applied insecticides on farm animals
c2pestfarm<-bc2$farmyrs[t19]

# Total years applied insecticides outside house
c2pestout<-bc2$outyrs[t19]

# Total years applied insecticides inside house
c2pestin<-bc2$inyrs[t19]

# Total years applied pesticides
c2pestpest<-bc2$pestyrs[t19]

### Binary Variables...###

# race (W=white,B=black)
c2race<-clue2$Race[t14]

# Gender (M=male,F=female)
c2gender<-clue2$Sex[t14]

# Exposed to pesticides,insecticides,herbicides,fungicides,or fumigants
on the job
# (0=no,1=yes)
c2occ<-bc2$pestcid[t19]

# current smoker (1=yes,2=n0) at time of blood draw
c2smoke<-clue2$NCig[t14]

# Are you a vegetarian now (avoid meat) (0=no,1=yes)
c2meat<-bc4$Vegnow[t20]

# Avoid eggs (0=no,1=yes)
c2eggs2<-bc4$Noeggs[t20]

# Avoid dairy products (0=no,1=yes)
c2dairy2<-bc4$Nodairy[t20]

### Categorical variables...###

# Marital status: 1=never married,2=married now,3=other,9=not stated
c2marital<-clue2$MS[t14]

#New egg variable from 89 diet study, how often did subject
# how often did subject eat eggs in last year
# 1=never,2=1-3/month,3=1/week,4=2-4/week,5=5-6/week,6=1/day
# 7=2-3/day,8=4-5/day,9=6+/day

c2eggs<-mixed$Egguse[t21]

# Same thing for fruit use in past year
#New egg variable from 89 diet study, how often did subject
# how often did subject eat eggs in last year
# 1=never,2=1-3/month,3=1/week,4=2-4/week,5=5-6/week,6=1/day
# 7=2-3/day,8=4-5/day,9=6+/day

c2fruit<-fruit$fruit[t22]

# Water source at current residence (0=municipal,1=household well
dug,2=well drilled
# 3=bottled, 4=cistern, 5=spring, 6=other)
c2water<-bc2$water1[t19]

# broiled/baked fish use in the last year
# (1=never,2=1-4/year,3=5-11/year,4=1-3/month,5=1/week,6=2-4/week,
# 7=almost everyday)

bakedfish<-diet$OfISHUSE[t15]
friedfish<-diet$Ffishuse[t15]
#c2fish<-diet$OfISHUSE[t15]+diet$Ffishuse[t15]
cor(diet$OfISHUSE,diet$Ffishuse, use="pairwise.complete.obs")

# Recode to make additions of these variables make sense
# 1 variable = (1=never,2=1-4/year,3=5-11/year,4=12-36/year,5=52/year,
# 6=104-208/year, 7=365/year
# Then take midpoints of these servings to get the average
serving/category
# 1 variable = (1=0 never,2=2.5/year,3=8/year,4=24/year,5=52/year,
# 6=156/year, 7=365/year

# Need to rename the variables to be these avg. servings/yr

baked1<-bakedfish
baked1[bakedfish==1]<-0
baked1[bakedfish==2]<-2.5
baked1[bakedfish==3]<-8
baked1[bakedfish==4]<-24
baked1[bakedfish==5]<-52
baked1[bakedfish==6]<-156
baked1[bakedfish==7]<-365

fried1<-friedfish
fried1[friedfish==1]<-0
fried1[friedfish==2]<-2.5
fried1[friedfish==3]<-8
fried1[friedfish==4]<-24
fried1[friedfish==5]<-52
fried1[friedfish==6]<-156
fried1[friedfish==7]<-365

fish1<-baked1+fried1
fish2<-fish1
fish2[fish1==2.5]<-1
```

```
fish2[fish1==5]<-2
fish2[fish1==8]<-3
fish2[fish1==10.5]<-4
fish2[fish1==16]<-5
fish2[fish1==24]<-6
fish2[fish1==26.5]<-7
fish2[fish1==32]<-8
fish2[fish1==48]<-9
fish2[fish1==52]<-10
fish2[fish1==54.5]<-11
fish2[fish1==60]<-12
fish2[fish1==76]<-13
fish2[fish1==104]<-14
fish2[fish1==156]<-15
fish2[fish1==158.5]<-16
fish2[fish1==164]<-17
fish2[fish1==180]<-18
fish2[fish1==208]<-19
fish2[fish1==312]<-20
fish2[fish1==365]<-21
fish2[fish1==367.5]<-22
fish2[fish1==389]<-23
fish2[fish1==730]<-24

# new coding for fish2 variable=how often eat fried/baked/broiled fish
in last year
# 0=never,1=2-3/year,2=5/yr,3=8/yr,4=10-11/yr,5=16/yr,6=2/month,

# new coding for c2fish variable=how often eat fried/baked/broiled fish
in last year
# 0=never,1=less than once a month,2= 1-2/month,3=1-2/week,4=3-4/week,
# 5=5-7/week,6=more than once a day

c2fish<-fish2
c2fish[fish2==1 - 4]<-1
c2fish[fish2==5 - 9]<-2
c2fish[fish2==10 -14]<-3
c2fish[fish2==15 -19]<-4
c2fish[fish2==20-21]<-5
c2fish[fish2==22-24]<-6

#New dairy variable from 89 diet study, coded as:
# how often did subject eat these dairy products in last year
# 1=never,2=1-3/month,3=1/week,4=2-4/week,5=5-6/week,6=1/day
# 7=2-3/day,8=4-5/day,9=6+/day

Cheese<-mixed$Cheeseuse[t21]
Whole<-mixed$WMilkuse[t21]
percent<-mixed$Milk2use[t21]
Skim<-mixed$Skimuse[t21]

#First need to recode these to make midpoint of serving #
# 1=0/month,2=2/month,3=4/month,4=12/month,5=22/month,6=30/month
#7=75/month,8=135/month,9=180/month

cheese1<-Cheese
cheese1[Cheese==1]<-0
```

```
cheese1[Cheese==2]<-2
cheese1[Cheese==3]<-4
cheese1[Cheese==4]<-12
cheese1[Cheese==5]<-22
cheese1[Cheese==6]<-30
cheese1[Cheese==7]<-75
cheese1[Cheese==8]<-135
cheese1[Cheese==9]<-180

whole1<-Whole
whole1[Whole==1]<-0
whole1[Whole==2]<-2
whole1[Whole==3]<-4
whole1[Whole==4]<-12
whole1[Whole==5]<-22
whole1[Whole==6]<-30
whole1[Whole==7]<-75
whole1[Whole==8]<-135
whole1[Whole==9]<-180

percent1<-percent
percent1[percent==1]<-0
percent1[percent==2]<-2
percent1[percent==3]<-4
percent1[percent==4]<-12
percent1[percent==5]<-22
percent1[percent==6]<-30
percent1[percent==7]<-75
percent1[percent==8]<-135
percent1[percent==9]<-180

skim1<-Skim
skim1[Skim==1]<-0
skim1[Skim==2]<-2
skim1[Skim==3]<-4
skim1[Skim==4]<-12
skim1[Skim==5]<-22
skim1[Skim==6]<-30
skim1[Skim==7]<-75
skim1[Skim==8]<-135
skim1[Skim==9]<-180

dairy1<-cheese1+whole1+percent1+skim1

# Recode variables to new categories
# Number of servings over past year
# 0=never,1=less than 1/week,2= 1-2/week,3=3-4/week,4=5-6/week,
# 5=everyday,6=1-2/day,7=3-4/day, 8=5 or more times a day

c2dairy<-dairy1
c2dairy[dairy1==2-11]<-1
c2dairy[dairy1==12-22]<-3
c2dairy[dairy1==23-27]<-4
c2dairy[dairy1==28-32]<-5
c2dairy[dairy1==33-78]<-6
c2dairy[dairy1==79-142]<-7
c2dairy[dairy1==143-210]<-8
```

```
#New vegetable variable from 89 diet study, coded as:
# how often did subject eat these veggies in last year
# 1=never,2=1-3/month,3=1/week,4=2-4/week,5=5-6/week,6=1/day
# 7=2-3/day,8=4-5/day,9=6+/day

salad<-diet$saluse[t15]
carrots<-diet$CARUSE[t15]
spinach<-diet$SPINUSE[t15]
yams<-diet$YAMUSE[t15]
potatoes<-diet$POTUSE[t15]
greens<-diet$GREENUSE[t15]
othervegs<-diet$OVEGUSE[t15]
brocoli<-diet$BROCUSE[t15]
cabbage<-diet$CABUSE[t15]

#First need to recode these to make midpoint of serving #
# 1=0/month,2=2/month,3=4/month,4=12/month,5=22/month,6=30/month
#7=75/month,8=135/month,9=180/month

salad1<-salad
salad1[salad==1]<-0
salad1[salad==2]<-2
salad1[salad==3]<-4
salad1[salad==4]<-12
salad1[salad==5]<-22
salad1[salad==6]<-30
salad1[salad==7]<-75
salad1[salad==8]<-135
salad1[salad==9]<-180

carrots1<-carrots
carrots1[carrots==1]<-0
carrots1[carrots==2]<-2
carrots1[carrots==3]<-4
carrots1[carrots==4]<-12
carrots1[carrots==5]<-22
carrots1[carrots==6]<-30
carrots1[carrots==7]<-75
carrots1[carrots==8]<-135
carrots1[carrots==9]<-180

yams1<-yams
yams1[yams==1]<-0
yams1[yams==2]<-2
yams1[yams==3]<-4
yams1[yams==4]<-12
yams1[yams==5]<-22
yams1[yams==6]<-30
yams1[yams==7]<-75
yams1[yams==8]<-135
yams1[yams==9]<-180

greens1<-greens
greens1[greens==1]<-0
greens1[greens==2]<-2
greens1[greens==3]<-4
```

```
greens1[greens==4]<-12
greens1[greens==5]<-22
greens1[greens==6]<-30
greens1[greens==7]<-75
greens1[greens==8]<-135
greens1[greens==9]<-180

cabbage1<-cabbage
cabbage1[cabbage==1]<-0
cabbage1[cabbage==2]<-2
cabbage1[cabbage==3]<-4
cabbage1[cabbage==4]<-12
cabbage1[cabbage==5]<-22
cabbage1[cabbage==6]<-30
cabbage1[cabbage==7]<-75
cabbage1[cabbage==8]<-135
cabbage1[cabbage==9]<-180

spinach1<-spinach
spinach1[spinach==1]<-0
spinach1[spinach==2]<-2
spinach1[spinach==3]<-4
spinach1[spinach==4]<-12
spinach1[spinach==5]<-22
spinach1[spinach==6]<-30
spinach1[spinach==7]<-75
spinach1[spinach==8]<-135
spinach1[spinach==9]<-180

potatoes1<-potatoes
potatoes1[potatoes==1]<-0
potatoes1[potatoes==2]<-2
potatoes1[potatoes==3]<-4
potatoes1[potatoes==4]<-12
potatoes1[potatoes==5]<-22
potatoes1[potatoes==6]<-30
potatoes1[potatoes==7]<-75
potatoes1[potatoes==8]<-135
potatoes1[potatoes==9]<-180

brocoli1<-brocoli
brocoli1[brocoli==1]<-0
brocoli1[brocoli==2]<-2
brocoli1[brocoli==3]<-4
brocoli1[brocoli==4]<-12
brocoli1[brocoli==5]<-22
brocoli1[brocoli==6]<-30
brocoli1[brocoli==7]<-75
brocoli1[brocoli==8]<-135
brocoli1[brocoli==9]<-180

othervegs1<-othervegs
othervegs1[othervegs==1]<-0
othervegs1[othervegs==2]<-2
othervegs1[othervegs==3]<-4
othervegs1[othervegs==4]<-12
othervegs1[othervegs==5]<-22
```

```
othervegs1[othervegs==6]<-30
othervegs1[othervegs==7]<-75
othervegs1[othervegs==8]<-135
othervegs1[othervegs==9]<-180

veggie1<-salad1+greens1+brocoli1+carrots1+yams1+potatoes1+othervegs1+
spinach1+cabbage1

# Recode variables to new categories
# Number of servings over past year
# 0=never,1=less than 1/week,2= 1-2/week,3=3-4/week,4=5-6/week,
# 5=everyday,6=1-2/day,7=3-4/day, 8=5 or more times a day

c2veggie<-veggie1
c2veggie[veggie1==2-5]<-1
c2veggie[veggie1==6-11]<-2
c2veggie[veggie1==12-22]<-3
c2veggie[veggie1==23-27]<-4
c2veggie[veggie1==28-32]<-5
c2veggie[veggie1==33-78]<-6
c2veggie[veggie1==79-142]<-7
c2veggie[veggie1==143-210]<-8

# to determine how many drinks/week currently when blood taken
# (1=never,2=1-3/month,3=1/week,4=2-4/week,5=5-6/week,6=1/day,7=2-
3/day,
# 8=4-5/day,9=6+per day
#c2alcohol<-diet$beeruse[t15]+diet$wINEcuse[t15]+diet$LIQURUSE[t15]

beer<-diet$beeruse[t15]
wine<-diet$wINEcuse[t15]
liquor<-diet$LIQURUSE[t15]

#First need to recode these to make midpoint of serving #
# 1=0/month,2=2/month,3=4/month,4=12/month,5=22/month,6=30/month
#7=75/month,8=135/month,9=180/month

beer1<-beer
beer1[beer==1]<-0
beer1[beer==2]<-2
beer1[beer==3]<-4
beer1[beer==4]<-12
beer1[beer==5]<-22
beer1[beer==6]<-30
beer1[beer==7]<-75
beer1[beer==8]<-135
beer1[beer==9]<-180

wine1<-wine
wine1[wine==1]<-0
wine1[wine==2]<-2
wine1[wine==3]<-4
wine1[wine==4]<-12
wine1[wine==5]<-22
wine1[wine==6]<-30
wine1[wine==7]<-75
wine1[wine==8]<-135
```

```
wine1[wine==9]<-180

liquor1<-liquor
liquor1[liquor==1]<-0
liquor1[liquor==2]<-2
liquor1[liquor==3]<-4
liquor1[liquor==4]<-12
liquor1[liquor==5]<-22
liquor1[liquor==6]<-30
liquor1[liquor==7]<-75
liquor1[liquor==8]<-135
liquor1[liquor==9]<-180

alcohol1<-beer1+wine1+liquor1

# Recode variables to new categories
# Number of servings over past year
# 0=never,1=less than 1/week,2= 1-2/week,3=3-4/week,4=5-6/week,
# 5=everyday,6=1-2/day,7=3-4/day, 8=5 or more times a day

c2alcohol<-alcohol1
c2alcohol[alcohol1==2-5]<-1
c2alcohol[alcohol1==6-11]<-2
c2alcohol[alcohol1==12-22]<-3
c2alcohol[alcohol1==23-27]<-4
c2alcohol[alcohol1==28-32]<-5
c2alcohol[alcohol1==33-78]<-6
c2alcohol[alcohol1==79-142]<-7
c2alcohol[alcohol1==143-210]<-8

# look at missing data using talbe(is.na( function (only shown for one)
table(is.na(c2dtosite))

#Check to make sure what these data look like, and make sure they are
continuous, binary, or categorical (one one shown):

stem(c2dtosite)

# make data frame including all covariates and outcome variables
c2data<-data.frame(clue2ID,clue1ID,c2addX,c2addY,c2dde,c2ddelip,
c2tpcbspeak, c2tpcbspeaklip,c22tpcbspeak,c22tpcbspeaklip,c2achlordane,
c2achlordanelip,c2gchlordane,c2gchlordanelip,c2oxychlor,c2oxychlorlip,
c2tnonachlor,c2tnonachlorlip,c2heptachlor,c2heptachlorlip,c2heptepox,
c2heptepoxlip,c2HCB,c2HCBlip,c2mirex,c2mirexlip,c2aldrin,c2aldrinlip,
c2bHCH,c2bHCHlip,c2gHCH,c2gHCHlip,c2zip,c2water,c2eggs2,c2dairy2,
c2veggie,tlp2,c2fruit,c2ddt,c2ddtlip,c2ddd,c2dddlip,c2dtosite,c2age,
c2edu,c2marital,c2BMI,c2race,c2gender,c2eggs,c2meat,c2dairy,c2fish,
c2occ,c2alcohol,c2smoke,c2brfeed,c2pestpets,c2pestfarm,c2pestout,
c2pestin,c2pestpest)
names(c2data)
dim(c2data)

# Need to make these categorical
c2data$c2fish<-as.factor(c2data$c2fish)
c2data$c2water<-as.factor(c2data$c2water)
c2data$c2eggs<-as.factor(c2data$c2eggs)
c2data$c2dairy<-as.factor(c2data$c2dairy)
```

```
c2data$c2alcohol<-as.factor(c2data$c2alcohol)
c2data$c2veggie<-as.factor(c2data$c2veggie)
c2data$c2fruit<-as.factor(c2data$c2fruit)
c2data$c2marital<-as.factor(c2data$c2marital)

# Check to make sure data matched up alright

c2data[c(113,356,512),]
oc2[c(15,349,678),]

# Now make some additional variables for possible analysis
# First is a few dtosite buffer variables

# make buffer in 5 mile radius from site
# (1=<5mi,2=5-10mi,3=10-15mi,4=15-20mi,5=20-25mi,6=25-30mi,7=>30mi)
x<-c2data$c2dtosite*5280
y<-c2data$c2dtosite*5280
x[y<=26400]<-1
x[y>26400&y<=52800]<-2
x[y>52800&y<=79200]<-3
x[y>79200&y<=105600]<-4
x[y>105600&y<=132000]<-5
x[y>132000&y<=158400]<-6
x[y>158400]<-7
c2data<-data.frame(c2data,c2dbuff2=x)

# and Make rural vs. urban variable

# get point of city origin
HagX<-1109576.10384
HagY<-720505.27649

# Make continuous variable as distance to city (c2dtocity) in miles

c2dtocity<-(sqrt((c2data$c2addX-HagX)^2+(c2data$c2addY-HagY)^2))/5280
c2data<-data.frame(c2data,c2dtocity=c2dtocity)

#Make Urban vs. Rural variable where rural is anything greater than
8000 ft (1.5miles) from
# center of city
x<-c2data$c2dtocity
y<-c2data$c2dtocity
x[y<=1.5]<-1
x[y>1.5]<-2

c2data<-data.frame(c2data,c2urban=x)

stem(c2data$c2dbuff2)
stem(c2data$c2dtocity)
stem(c2data$c2urban)

sitetocity<-(sqrt((c2data$sitex-HagX)^2+(c2data$sity-HagY)^2))/5280

# Make a direction variable to see if direction around site makes a
difference, and rotate every 15 degrees using DirBins2 defined as:

DirBins2<-function(data,rot=0){
  dsite<-data$c2dtosite*5280
  x<-data$c2addX
  y<-data$c2addY
  dx<-x-sitex
  dy<-y-sitey
  n<-length(x)
  quad<-numeric(n)
  quad[dx>0&dy>0]<-1
  quad[dx>0&dy<0]<-2
  quad[dx<0&dy<0]<-3
  quad[dx<0&dy>0]<-4
  angle<-rep(NA,n)
  angle[quad==1]<-asin(abs(dx[quad==1])/dsite[quad==1])
  angle[quad==2]<-90*pi/180 + asin(abs(dy[quad==2])/dsite[quad==2])
  angle[quad==3]<-pi + asin(abs(dx[quad==3])/dsite[quad==3])
  angle[quad==4]<-270*pi/180 + asin(abs(dy[quad==4])/dsite[quad==4])
  angle<-angle*180/pi
  angle<-angle+rot
  bin<-rep(NA,n)
  bin[angle>0 & angle <= 45]<-1
  bin[angle>45 & angle <= 90]<-2
  bin[angle>90 & angle <= 135]<-3
  bin[angle>135 & angle <= 180]<-4
  bin[angle>180 & angle <= 225]<-5
  bin[angle>225 & angle <= 270]<-6
  bin[angle>270 & angle <= 315]<-7
  bin[angle>315 & angle <= 360]<-8
  return(bin)
}

c2dbins0<-DirBins2(c2data)
c2data<-data.frame(c2data,c2dbins0=c2dbins0)
cor(c2data$c2dde[c2data$c2dbins0==1],c2data$c2dtosite[c2data$c2dbins0==
1],use="pairwise.complete.obs")
for(i in 1:8)
print(cor(c2data$c2dde[c2data$c2dbins0==i],c2data$c2dtosite[c2data$c2db
ins0==i],use="pairwise.complete.obs"))

c2dbins15<-DirBins2(c2data,rot=15)
c2data<-data.frame(c2data,c2dbins15=c2dbins15)
for(i in 1:8)
print(cor(c2data$c2dde[c2data$c2dbins15==i],c2data$c2dtosite[c2data$c2d
bins15==i],use="pairwise.complete.obs"))

c2dbins30<-DirBins2(c2data,rot=30)
c2data<-data.frame(c2data,c2dbins30=c2dbins30)
for(i in 1:8)
print(cor(c2data$c2dde[c2data$c2dbins30==i],c2data$c2dtosite[c2data$c2d
bins30==i],use="pairwise.complete.obs"))

# Make those that need to be categorical, categorical
c2data$c2dbuff2<-as.factor(c2data$c2dbuff2)
c2data$c2dbins0<-as.factor(c2data$c2dbins0)
c2data$c2dbins15<-as.factor(c2data$c2dbins15)
c2data$c2dbins30<-as.factor(c2data$c2dbins30)

# Now there are 2 people in clue 2 that participated in 2 studies.
# same person, same blood, same location
# I will average their outcome levels.

#1
c(1:797)[c2data$clue2ID=="B-4761"]
c2data[c2data$clue2ID=="B-4761",]
c2data$clue1ID[283]<-"5047a"
c2data$clue1ID[284]<-"5047b"
c2data[c(283,284),]

c2data$clue1ID[c2data$clue1ID=="5047b"]<-"5047"

u<-c2data$c2dde[283]
v<-c2data$c2dde[284]
c2data$c2dde[c2data$clue1ID=="5047"]<-mean(c(u,v))
u<-c2data$c2ddelip[283]
v<-c2data$c2ddelip[284]
c2data$c2ddelip[c2data$clue1ID=="5047"]<-mean(c(u,v))

a<-c(1:797)[c2data$clue1ID=="5047a"]
c2data<-c2data[-a,]
c2data[c2data$clue2ID=="B-4761",]

#2
c(1:796)[c2data$clue2ID=="B-6568"]
c2data[c2data$clue2ID=="B-6568",]
c2data$clue1ID[315]<-"a"
c2data$clue1ID[316]<-"b"
c2data[315:316,]

u<-c2data$c2dde[315]
v<-c2data$c2dde[316]
c2data$c2dde[c2data$clue1ID=="b"]<-mean(c(u,v))
c2data$c2ddelip[c2data$clue1ID=="b"]<-
mean(c(u,v))*c2data$tlp2[c2data$clue1ID=="a"]

u<-c2data$c22tpcbspeak[315]
v<-c2data$c22tpcbspeak[316]
c2data$c22tpcbspeak[c2data$clue1ID=="b"]<-mean(c(u,v))
c2data$c22tpcbspeaklip[c2data$clue1ID=="b"]<-
mean(c(u,v))*c2data$tlp2[c2data$clue1ID=="a"]

v<-c2data$c2achlordane[316]
c2data$c2achlordanelip[c2data$clue1ID=="b"]<-
v*c2data$tlp2[c2data$clue1ID=="a"]

v<-c2data$c2gchlordane[316]
c2data$c2gchlordanelip[c2data$clue1ID=="b"]<-
v*c2data$tlp2[c2data$clue1ID=="a"]

v<-c2data$c2oxychlor[316]
c2data$c2oxychlorlip[c2data$clue1ID=="b"]<-
v*c2data$tlp2[c2data$clue1ID=="a"]

v<-c2data$c2tnonachlor[316]
c2data$c2tnonachlorlip[c2data$clue1ID=="b"]<-
v*c2data$tlp2[c2data$clue1ID=="a"]
```

```
v<-c2data$c2HCB[316]
c2data$c2HCBlip[c2data$clue1ID=="b"]<-
v*c2data$tlp2[c2data$clue1ID=="a"]

v<-c2data$c2heptachlor[316]
c2data$c2heptachlorlip[c2data$clue1ID=="b"]<-
v*c2data$tlp2[c2data$clue1ID=="a"]

v<-c2data$c2hepepox[316]
c2data$c2hepepoxlip[c2data$clue1ID=="b"]<-
v*c2data$tlp2[c2data$clue1ID=="a"]

v<-c2data$c2mirex[316]
c2data$c2mirexlip[c2data$clue1ID=="b"]<-
v*c2data$tlp2[c2data$clue1ID=="a"]

v<-c2data$c2aldrin[316]
c2data$c2aldrinlip[c2data$clue1ID=="b"]<-
v*c2data$tlp2[c2data$clue1ID=="a"]

v<-c2data$c2dieldrin[316]
c2data$c2dieldrinlip[c2data$clue1ID=="b"]<-
v*c2data$tlp2[c2data$clue1ID=="a"]

v<-c2data$c2endrin[316]
c2data$c2endrinlip[c2data$clue1ID=="b"]<-
v*c2data$tlp2[c2data$clue1ID=="a"]

v<-c2data$c2bHCH[316]
c2data$c2bHCHlip[c2data$clue1ID=="b"]<-
v*c2data$tlp2[c2data$clue1ID=="a"]

v<-c2data$c2gHCH[316]
c2data$c2gHCHlip[c2data$clue1ID=="b"]<-
v*c2data$tlp2[c2data$clue1ID=="a"]

v<-c2data$c2ddt[316]
c2data$c2ddtlip[c2data$clue1ID=="b"]<-
v*c2data$tlp2[c2data$clue1ID=="a"]

v<-c2data$c2ddd[316]
c2data$c2dddlip[c2data$clue1ID=="b"]<-
v*c2data$tlp2[c2data$clue1ID=="a"]

c2data[315:316,]
dim(c2data)

c2data$clue1ID[c2data$clue1ID=="b"]<-"<NA>"

b<-c(1:796)[c2data$clue1ID=="a"]
c2data<-c2data[-b,]
c2data[c2data$clue2ID=="B-6568",]

# Take out variables that aren't in the WC zip codes
u<-c(1:795)[c2data$c2zip==21736|c2data$c2zip==17225|
c2data$c2zip==22993|c2data$c2zip==21741]
```

```
c2data<-c2data[-u,]

# Take out 3 points where geocoded location is not inside washington
County
dim(c2data)
v<-c(1:790)[c2data$clue2ID=="L-1838"|c2data$clue2ID=="A-
1188"|c2data$clue2ID=="M-3492"]
c2data<-c2data[-v,]

# Descriptive statistics on outcome variables

Mc2dde<-mean.default(c2data$c2dde,na.rm=T)
Nc2dde<-length(c2data$c2dde[!is.na(c2data$c2dde)])
SDc2dde<-sqrt(var(c2data$c2dde,na.rm=T))/sqrt(Nc2dde)
CILc2dde<-Mc2dde-1.96*SDc2dde
CIUc2dde<-Mc2dde+1.96*SDc2dde
Mc2dde
Nc2dde
SDc2dde
CILc2dde
CIUc2dde
summary(c2data$c2dde)

# First, clean up data (NAs in coordinates or outcome variable)
# And make them into geodata objects for Variogram analyses for OC data

# making dataframe

c2geodata<-data.frame(c2addX=c2data$c2addX,c2addY=c2data$c2addY,c2dde=
c2data$c2dde,c2ddelip=c2data$c2ddelip,c22tpcbspeak=c2data$c22tpcbspeak,
c22tpcbspeaklip=c2data$c22tpcbspeaklip,clue2ID=c2data$clue2ID,
c2achlordane=c2data$c2achlordane,c2achlordanelip=
c2data$c2achlordanelip,c2gchlordane=c2data$c2gchlordane,c2gchlordanelip
=c2data$c2gchlordanelip,c2oxychlor=c2data$c2oxychlor,c2oxychlorlip=
c2data$c2oxychlorlip,c2tnonachlor=c2data$c2tnonachlor,c2tnonachlorlip=
c2data$c2tnonachlorlip,c2heptachlor=c2data$c2heptachlor,c2heptachlorlip
=c2data$c2heptachlorlip,c2heptepox=c2data$c2heptepox,c2heptepoxlip=
c2data$c2heptepoxlip,c2HCB=c2data$c2HCB,c2HCBlip=c2data$c2HCBlip,c2mire
x=c2data$c2mirex,c2mirexlip=c2data$c2mirexlip,c2aldrin=c2data$c2aldrin,
c2aldrinlip=c2data$c2aldrinlip,c2bHCH=c2data$c2bHCH,c2bHCHlip=
c2data$c2bHCHlip,c2gHCH=c2data$c2gHCH,c2gHCHlip=c2data$c2gHCHlip,c2ddt=
c2data$c2ddt,c2ddtlip=c2data$c2ddtlip,c2ddd=c2data$c2ddd,c2dddlip=
c2data$c2dddlip)

# Taking out missing data for DDE
dim(c2geodata)
u1<-c(1:787)[is.na(c2geodata$c2dde)]
u2<-c(1:787)[c2geodata$c2dde<=0]
u3<-c(1:787)[is.na(c2geodata$c2addX)|is.na(c2geodata$c2addY)]
u<-c(u1,u2,u3)
u<-unique(u)

# Make geodata objects without the missing data

c2ddegeo<-as.geodata(c2geodata[-u,],data.col=3,covar.col=7)
```

```
# Write these into text files to bring them into GIS: R->text->excel-
>dbf->GIS

write.table(c2geodata[-u,],file="c:\\shenshaw\\Thesis\\CLUE\\CLUE_Data
\\Changed_Variable_Names\\Merged_Data\\c2ddegeodata.csv",sep=",",col.na
mes=T,row.names=F)

## Exaustive Regression Search for only DDE ##########################

library("combinat")

# Only one search for one compound is shown here to save space, but
this was done for all organochlorines, both lipid adjusted and
unadjusted

Furthermore, only one full loop is shown, and a description of the
additional loops run for all compounds are described at the end of the
loop.

###DDE###

# Make new dataset with only those variables I wish to test together

c2dde89a<-c2data[,c("c2dde","c2dtosite","c2age","c2edu","c2marital",
"c2race","c2veggie","c2fruit","c2gender","c2BMI","c2smoke","c2fish",
"c2dairy","c2eggs","c2alcohol","c2dbins0")]

# Have to run this code each time and change data name

RunReg<-function(vars,resp="c2dde"){
  xnam<-names(c2dde89a)[vars]
  fmla<-as.formula(paste(resp," ~ ", paste(xnam,collapse=" + ")))
  fit<-lm(fmla,data=c2dde89a)
  rsqrd<-round(100*summary(fit)$adj.r.squared,2)
  fmla<-as.character(fmla)
  fmla<-paste(fmla[2],paste(" ",fmla[1]," ",sep=""),fmla[3],sep="")
  return(c(rsqrd,as.character(fmla)))
}

# This has 16 possible variables and one response listed in the first
column. Below you'll need t1 through tN, where N is the number of
possible variables

models<-combn(2:16,1)
t1<-apply(models,2,RunReg)

models<-combn(2:16,2)
t2<-apply(models,2,RunReg)

models<-combn(2:16,3)
t3<-apply(models,2,RunReg)

models<-combn(2:16,4)
t4<-apply(models,2,RunReg)

models<-combn(2:16,5)
t5<-apply(models,2,RunReg)
```

```
models<-combn(2:16,6)
t6<-apply(models,2,RunReg)

models<-combn(2:16,7)
t7<-apply(models,2,RunReg)

models<-combn(2:16,8)
t8<-apply(models,2,RunReg)

models<-combn(2:16,9)
t9<-apply(models,2,RunReg)

models<-combn(2:16,10)
t10<-apply(models,2,RunReg)

models<-combn(2:16,11)
t11<-apply(models,2,RunReg)

models<-combn(2:16,12)
t11a<-apply(models,2,RunReg)

models<-combn(2:16,13)
t11b<-apply(models,2,RunReg)

models<-combn(2:16,14)
t11c<-apply(models,2,RunReg)

models<-matrix(combn(2:16,15),ncol=1)
t12<-apply(models,2,RunReg)

tall<-cbind(t1,t2,t3,t4,t5,t6,t7,t8,t9,t10,t11,t11a,t11b,t11c,t12)
Results<-data.frame(t(tall))
names(Results)<-c("AdjRsqrd","Model")
Results$AdjRsqrd<-as.numeric(as.vector(Results$AdjRsqrd))
u<-order(Results$AdjRsqrd,decreasing=TRUE)
Results[u[1:10],]

# b Try using addX,addY instead of dtosite

c2dde89b<-c2data[,c("c2dde","c2addX","c2addY","c2age","c2edu",
"c2marital","c2race","c2veggie","c2fruit","c2gender","c2BMI","c2smoke",
"c2fish", "c2dairy","c2eggs","c2alcohol","c2dbins0")]

#c Try using distance buffer instead of distance to site/addX/addY

c2dde89c<-c2data[,c("c2dde","c2dbuff2","c2age","c2edu","c2marital",
"c2race","c2veggie","c2fruit","c2gender","c2BMI","c2smoke","c2fish",
"c2dairy","c2eggs","c2alcohol","c2dbins0")]

#d Try using urban/rural variable instead of other distance to site
variables

c2dde89d<-c2data[,c("c2dde","c2urban","c2age","c2edu","c2marital",
"c2race","c2veggie","c2fruit","c2gender","c2BMI","c2smoke","c2fish","c2
dairy", "c2eggs","c2alcohol","c2dbins0")]
```

```
# Now do steps a-d after rotating direction bin 15 degrees

c2dde89a15<-c2data[,c("c2dde","c2dtosite","c2age","c2edu","c2marital",
"c2race","c2veggie","c2fruit","c2gender","c2BMI","c2smoke", "c2fish",
"c2dairy","c2eggs","c2alcohol","c2dbins15")]

# Now do steps a-d after rotating direction bin 30 degrees

c2dde89a30<-c2data[,c("c2dde","c2dtosite","c2age","c2edu","c2marital",
"c2race","c2veggie","c2fruit","c2gender","c2BMI","c2smoke","c2fish",
"c2dairy","c2eggs","c2alcohol","c2dbins30")]

# Now get the results of all of these combinations and order according
to R2:

tall<-cbind(t1,t2,t3,t4,t5,t6,t7,t8,t9,t10,t11,t11a,t11b,t11c,t12,t14,
t15,t16,t17,t18,t19,t20,t21,t22,t23,t24,t25,t25a,t25b,t25c,t26,t28,t29,
t30,t31,t32,t33,t34,t35,t36,t37,t38,t38a,t38b,t38c,t39,t41,t42,t43,t44,
t45,t46,t47,t48,t49,t50,t51,t51a,t51a,t51b,t51c,t52,t101,t102,t103,t104
,t105,t106,t107,t108,t109,t110,t111,t111a,t111b,t111c,t112,t114,t115,
t116,t117,t118,t119,t120,t121,t122,t123,t124,t125,t125a,t125b,t125c,
t126,t128,t129,t130,t131,t132,t133,t134,t135,t136,t137,t138,t138a,
t138b,t138c,t139,t141,t142,t143,t144,t145,t146,t147,t148,t149,t150,
t151,t151a,t151b,t151c,t152,t201,t202,t203,t204,t205,t206,t207,t208,
t209,t210,t211,t211a,t211b,t211c,t212,t214,t215,t216,t217,t218,t219,
t220,t221,t222,t223,t224,t225,t225a,t225b,t225c,t226,t228,t229,t230,
t231,t232,t233,t234,t235,t236,t237,t238,t238a,t238b,t238c,t239,t241,
t242,t243,t244,t245,t246,t247,t248,t249,t250,t251,t251a,t251b,t251c,
t252)
Results<-data.frame(t(tall))
names(Results)<-c("AdjRsqrd","Model")
Results$AdjRsqrd<-as.numeric(as.vector(Results$AdjRsqrd))
u<-order(Results$AdjRsqrd,decreasing=TRUE)
Results[u[1:40],]

write(t(Results[u,]),file="c:\\shenshaw\\Thesis\\Papers\\Paper3\\dde89.
txt",ncol=2)

# Now models are Chosen:

# use whites only, although all analysis was also performed with entire
population as well.  Whites only are reported in Manuscript Two

c2data3<-c2data[c2data$c2race2==1,]

# try recoding dairy, veggie, and alcohol to only 3 categories
# Recode variables to new categories
# Old category: Number of servings over past year
# 0=never,1=less than 1/week,2= 1-2/week,3=3-4/week,4=5-6/week,
# 5=everyday,6=1-2/day,7=3-4/day, 8=5 or more times a day
# New category:
# 0=never, 1=seldom (less than once a week), 2=occasionally(less than
once a day),
# 3=often (everyday or more)

n<-length(c2data3$c2dairy)
c2dairy2<-rep(NA,n)
```

```
c2data3<-data.frame(c2data3,c2dairy2=c2dairy2)
c2data3$c2dairy2[c2data3$c2dairy==0]<-0
c2data3$c2dairy2[c2data3$c2dairy==1]<-1
c2data3$c2dairy2[c2data3$c2dairy==2|c2data3$c2dairy==3|c2data3$c2dairy=
=4]<-2
c2data3$c2dairy2[c2data3$c2dairy==5|c2data3$c2dairy==6|c2data3$c2dairy=
=7|c2data3$c2dairy==8]<-3
c2dairy2<-c2data3$c2dairy2
table(c2data3$c2dairy2)

# do the same with veggies and alcohol

n<-length(c2data3$c2veggie)
c2veggie2<-rep(NA,n)
c2data3<-data.frame(c2data3,c2veggie2=c2veggie2)
c2data3$c2veggie2[c2data3$c2veggie==0]<-0
c2data3$c2veggie2[c2data3$c2veggie==1]<-1
c2data3$c2veggie2[c2data3$c2veggie==2|c2data3$c2veggie==3|c2data3$c2veg
gie==4]<-2
c2data3$c2veggie2[c2data3$c2veggie==5|c2data3$c2veggie==6|c2data3$c2veg
gie==7|c2data3$c2veggie==8]<-3
c2veggie2<-c2data3$c2veggie2
table(c2data3$c2veggie2)

n<-length(c2data3$c2alcohol)
c2alcohol2<-rep(NA,n)
c2data3<-data.frame(c2data3,c2alcohol2=c2alcohol2)
c2data3$c2alcohol2[c2data3$c2alcohol==0]<-0
c2data3$c2alcohol2[c2data3$c2alcohol==1]<-1
c2data3$c2alcohol2[c2data3$c2alcohol==2|c2data3$c2alcohol==3|c2data3$c2
alcohol==4]<-2
c2data3$c2alcohol2[c2data3$c2alcohol==5|c2data3$c2alcohol==6|c2data3$c2
alcohol==7|c2data3$c2alcohol==8]<-3
c2alcohol2<-c2data3$c2alcohol2
table(c2data3$c2alcohol2)

n<-length(c2data3$c2race)
c2race2<-rep(NA,n)
c2data3<-data.frame(c2data3,c2race2=c2race2)
c2data3$c2race2[c2data3$c2race=="B"]<-2
c2data3$c2race2[c2data3$c2race=="W"]<-1

c2race2<-c2data3$c2race2
table(c2data3$c2race2)

c2data3$c2dairy2<-as.factor(c2data3$c2dairy2)
c2data3$c2veggie2<-as.factor(c2data3$c2veggie2)
c2data3$c2alcohol2<-as.factor(c2data3$c2alcohol2)

# Models chosen

c2ddereg<-lm(c2dde ~ c2age + c2edu  + c2veggie2 + c2gender + c2BMI +
c2smoke + c2dairy2,c2data3)
summary(c2ddereg)

# Analyzing models
# already made geodata objects and some analysis from Rc2
```

```
plot(variog(c2ddegeo,max.dist=100000))
title(" variogram for DDE 89")
plot(variog4(c2ddegeo,max.dist=100000))
title("multidirectional variogram for DDE 89")

# logged

plot(variog(c2ddegeo,max.dist=100000,lambda=0))
title(" variogram for logged DDE 89")
plot(variog4(c2ddegeo,max.dist=100000,lambda=0))
title("multidirectional variogram for logged DDE 89")

# fit variograms to residuals of models

u<-complete.cases(c2data3[,c("c2dde","c2age","c2edu","c2race2",
"c2gender","c2veggie2","c2smoke","c2BMI","c2dairy2")])
c2ddereggeo<-as.geodata(cbind(c2data3[u,c("c2addX","c2addY")],
as.vector(c2ddereg$residuals)))
plot(variog(c2ddereggeo,max.dist=100000))
title("V residuals of dde 89 model")
plot(variog4(c2ddereggeo,max.dist=100000))
title("Multidirectional V residuals of dde 89 model")

# check to make sure lined up alright
t1<-c2data3[u,c("c2addX","c2addY")]
t2<-c2ddereg$residuals
range(as.numeric(labels(t1)[[1]])-as.numeric(labels(t2)))

# plot residuals of models
plot(c2ddereg$residuals)
title("dde 89 residuals")

# make tables and summaries of categorical variables to see what is
# going on (do this for all covariates, although only gender is shown)

summary(c2data3$c2dtosite)

table(c2data3$c2gender,c2data3$c2gender)
table(c2data3$c2gender,c2data3$c2water)
table(c2data3$c2gender,c2data3$c2marital)
table(c2data3$c2gender,c2data3$c2fish)
table(c2data3$c2gender,c2data3$c2eggs)
table(c2data3$c2gender,c2data3$c2dairy2)
table(c2data3$c2gender,c2data3$c2veggie2)
table(c2data3$c2gender,c2data3$c2fruit)
table(c2data3$c2gender,c2data3$c2alcohol2)
table(c2data3$c2gender,c2data3$c2smoke)

cor(c2data3[,34:73], use="pairwise.complete.obs")
```

Now GLS regression is completed to take into account spatial dependence in the data. Two iterations are completed to assure best fit model.

```
# Producing GLS regression correction results CLUE 2

######### DDE ##################################
```

```
u<-complete.cases(c2data3[,c("c2dde","c2addX","c2addY","c2age",
"c2edu","c2veggie2","c2gender","c2BMI","c2smoke","c2dairy2")])
c2tempdde<-c2data3[u,]
names(c2tempdde)
c2covarsdde<-model.matrix(~c2age+c2edu+c2veggie2+c2gender+
c2BMI+c2smoke+c2dairy2,data=c2tempdde)
c2covarsdde<-data.frame(c2covarsdde)
c2covarsdde<-c2covarsdde[,-1]
# 5th column is dde, x, y are 3,4
c2temp2dde<-data.frame(c2tempdde[,c(3:4,5)],c2covarsdde)
c2t2geodde<-as.geodata(c2temp2dde,coords.col=1:2,data.col=3,
covar.col=4:12)

c2vg2adde<-variog(c2t2geodde,max.dist=100000)
plot(c2vg2adde)

c2trend1dde<-formula(~c2t2geodde$covariate$c2age+
c2t2geodde$covariate$c2edu+c2t2geodde$covariate$c2veggie23+
c2t2geodde$covariate$c2genderM+c2t2geodde$covariate$c2BMI+
c2t2geodde$covariate$c2smoke+c2t2geodde$covariate$c2dairy21+
c2t2geodde$covariate$c2dairy22+c2t2geodde$covariate$c2dairy23)

c2vg2bdde<-variog(c2t2geodde,max.dist=100000,trend=c2trend1dde)
plot(c2vg2bdde)
c2vg2bdde$beta.ols

# check to make sure these match the betas for lm method

ddelmc2temp2<-lm(c2dde~c2age+c2edu+c2veggie2+c2gender+c2BMI+c2smoke+
                      c2dairy2,data=c2tempdde)
ddelmc2temp2$coefficients

c2locs1dde<-c2t2geodde$cords

# Estimate parameters from residual variation in OLS model

cov1<-c(14,30000)
n<-dim(c2tempdde)[1]
nugget1<-28

fic2t2bdde<-likfit(c2t2geodde, cov.model="exp", ini.cov.pars=cov1,
method="reml",nugget=nugget1, trend=c2trend1dde)

plot(c2vg2bdde)
lines.variomodel(fic2t2bdde)
# beta GLS coef. and GLS std error
c2glsddea<-cbind(fic2t2bdde$beta,sqrt(diag(fic2t2bdde$beta.var)),
fic2t2bdde[[7]]/(sqrt(diag(fic2t2bdde[[8]]))))
c2glsddea

# get R2 and adjusted R2

dmatc2dde<-as.matrix(dist(c2temp2dde[,1:2]))
Cov<-
cov.spatial(dmatc2dde,cov.model="exp",cov.pars=fic2t2bdde$cov.pars)
```

```
Cov<-Cov/fic2t2bdde$cov.pars[1] # so Now we have (sigma^2)V as
covariance structure
Cinv<-solve(Cov)
X<-as.matrix(cbind(1,c2temp2dde[,4:12]))
p<-dim(X)[2]
Y<-matrix(c2temp2dde[,3],ncol=1)
beta.ols<-solve(t(X)%*%X)%*%t(X)%*%Y
var.ols<-diag(solve(t(X)%*%X))
sse.ols<-t((Y-X%*%beta.ols))%*%(Y-X%*%beta.ols)
mse.ols<-sse.ols/(n-p)
ols<-cbind(beta.ols,sqrt(mse.ols*var.ols))
sqrt(mse.ols)

# matches residual standard error from summary(temp2)
summary(c2temp2dde)

beta.gls<-solve(t(X)%*%Cinv%*%X)%*%t(X)%*%Cinv%*%Y
var.gls<-diag(solve(t(X)%*%Cinv%*%X))
sse.gls<-t((Y-X%*%beta.gls))%*%Cinv%*%(Y-X%*%beta.gls)
mse.gls<-sse.gls/(n-p)
gls<-cbind(beta.gls,sqrt(mse.gls*var.gls))

J<-matrix(1,n,1)
ssto.gls<-t(Y)%*%Cinv%*%Y-
solve(t(J)%*%Cinv%*%J)%*%(t(Y)%*%Cinv%*%J%*%t(J)%*%Cinv%*%Y)
ssr.gls<-ssto.gls-sse.gls
R2.gls<-1-sse.gls/ssto.gls
R2adj.gls<-1-((n-1)/(n-p))*sse.gls/ssto.gls

e.ols<-Y-X%*%beta.ols
e.gls<-Y-X%*%beta.gls

muhat<-X%*%beta.gls
resids<-Y-muhat

c2ddegeo2<-as.geodata(cbind(c2t2geodde$coords,resids))

plot(variog(c2ddegeo2))
plot(variog(c2ddegeo2, max.dist=100000))

# re-estimate initial parameters using residual variation of GLS model
(2nd iteration)

cov1<-c(180,30000)
nugget1<-80

fic2t2bdde2<-likfit(c2t2geodde, cov.model="exp", ini.cov.pars=cov1,
method="reml",nugget=nugget1, trend=c2trend1dde)

c2glsddeb<-cbind(fic2t2bdde2$beta,sqrt(diag(fic2t2bdde2$beta.var)),
fic2t2bdde2[[7]]/(sqrt(diag(fic2t2bdde2[[8]]))))
c2glsddeb

dmatc2dde<-as.matrix(dist(c2temp2dde[,1:2]))
Cov<-
cov.spatial(dmatc2dde,cov.model="exp",cov.pars=fic2t2bdde2$cov.pars)
```

```
Cov<-Cov/fic2t2bdde2$cov.pars[1] # so Now we have (sigma^2)V as
covariance structure
Cinv<-solve(Cov)
X<-as.matrix(cbind(1,c2temp2dde[,4:12]))
p<-dim(X)[2]
Y<-matrix(c2temp2dde[,3],ncol=1)
beta.ols<-solve(t(X)%*%X)%*%t(X)%*%Y
var.ols<-diag(solve(t(X)%*%X))
sse.ols<-t((Y-X%*%beta.ols))%*%(Y-X%*%beta.ols)
mse.ols<-sse.ols/(n-p)
ols<-cbind(beta.ols,sqrt(mse.ols*var.ols))
sqrt(mse.ols)

# matches residual standard error from summary(temp2)
summary(c2temp2dde)

beta.gls<-solve(t(X)%*%Cinv%*%X)%*%t(X)%*%Cinv%*%Y
var.gls<-diag(solve(t(X)%*%Cinv%*%X))
sse.gls<-t((Y-X%*%beta.gls))%*%Cinv%*%(Y-X%*%beta.gls)
mse.gls<-sse.gls/(n-p)
gls<-cbind(beta.gls,sqrt(mse.gls*var.gls))

J<-matrix(1,n,1)
ssto.gls<-t(Y)%*%Cinv%*%Y-
solve(t(J)%*%Cinv%*%J)%*%(t(Y)%*%Cinv%*%J%*%t(J)%*%Cinv%*%Y)
ssr.gls<-ssto.gls-sse.gls
R2.gls2<-1-sse.gls/ssto.gls
R2adj.gls2<-1-((n-1)/(n-p))*sse.gls/ssto.gls

c2glsdderesults1<-cbind(R2.gls, R2adj.gls, R2.gls2, R2adj.gls2)
c2glsdderesults1
c2glsdderesults<-cbind(c2glsddea, c2glsddeb)
c2glsdderesults

# simple (crude) univariate GLS regressions

fic2t2bdde2<-likfit(c2t2geodde, cov.model="exp", ini.cov.pars=cov1,
method="reml",nugget=nugget1, trend=~c2t2geodde$covariate$c2age)
c2glsdde2<-c(fic2t2bdde2$beta,sqrt(diag(fic2t2bdde2$beta.var)),
fic2t2bdde2$beta/(sqrt(diag(fic2t2bdde2$beta.var))))

fic2t2bdde3<-likfit(c2t2geodde, cov.model="exp", ini.cov.pars=cov1,
method="reml",nugget=nugget1, trend=~c2t2geodde$covariate$c2genderM)
c2glsdde3<-c(fic2t2bdde3$beta,sqrt(diag(fic2t2bdde3$beta.var)),
fic2t2bdde3$beta/(sqrt(diag(fic2t2bdde3$beta.var))))

fic2t2bdde5<-likfit(c2t2geodde, cov.model="exp", ini.cov.pars=cov1,
method="reml",nugget=nugget1, trend=~c2t2geodde$covariate$c2edu)
c2glsdde5<-c(fic2t2bdde5$beta,sqrt(diag(fic2t2bdde5$beta.var)),
fic2t2bdde5$beta/(sqrt(diag(fic2t2bdde5$beta.var))))

fic2t2bdde6<-likfit(c2t2geodde, cov.model="exp", ini.cov.pars=cov1,
method="reml",nugget=nugget1, trend=~c2t2geodde$covariate$c2smoke)
c2glsdde6<-c(fic2t2bdde6$beta,sqrt(diag(fic2t2bdde6$beta.var)),
fic2t2bdde6$beta/(sqrt(diag(fic2t2bdde6$beta.var))))
```

```
fic2t2bdde7<-likfit(c2t2geodde, cov.model="exp", ini.cov.pars=cov1,
method="reml", nugget=nugget1, trend=~c2t2geodde$covariate$c2dairy21)
c2glsdde7<-c(fic2t2bdde7$beta,sqrt(diag(fic2t2bdde7$beta.var)),
fic2t2bdde7$beta/(sqrt(diag(fic2t2bdde7$beta.var))))

fic2t2bdde8<-likfit(c2t2geodde, cov.model="exp", ini.cov.pars=cov1,
method="reml",nugget=nugget1, trend=~c2t2geodde$covariate$c2dairy22)
c2glsdde8<-c(fic2t2bdde8$beta,sqrt(diag(fic2t2bdde8$beta.var)),
fic2t2bdde8$beta/(sqrt(diag(fic2t2bdde8$beta.var))))

fic2t2bdde9<-likfit(c2t2geodde, cov.model="exp", ini.cov.pars=cov1,
method="reml",nugget=nugget1, trend=~c2t2geodde$covariate$c2dairy23)
c2glsdde9<-c(fic2t2bdde9$beta,sqrt(diag(fic2t2bdde9$beta.var)),
fic2t2bdde9$beta/(sqrt(diag(fic2t2bdde9$beta.var))))

fic2t2bdde10<-likfit(c2t2geodde, cov.model="exp", ini.cov.pars=cov1,
method="reml",nugget=nugget1, trend=~c2t2geodde$covariate$c2veggie23)
c2glsdde10<-c(fic2t2bdde10$beta,sqrt(diag(fic2t2bdde10$beta.var)),
fic2t2bdde10$beta/(sqrt(diag(fic2t2bdde10$beta.var))))

fic2t2bdde11<-likfit(c2t2geodde, cov.model="exp", ini.cov.pars=cov1,
method="reml",nugget=nugget1, trend=~c2t2geodde$covariate$c2BMI)
c2glsdde11<-c(fic2t2bdde11$beta,sqrt(diag(fic2t2bdde11$beta.var)),
fic2t2bdde11$beta/(sqrt(diag(fic2t2bdde11$beta.var))))

simresultsc2dde<-rbind(c2glsdde2,c2glsdde3,c2glsdde5,c2glsdde6,
c2glsdde7,c2glsdde8,c2glsdde9,c2glsdde10,c2glsdde11)

write.table(simresultsc2dde,file="c:\\shenshaw\\Thesis\\Papers\\Paper3\
\gls\\simc2dde.csv",sep=",",col.names=T,row.names=F)

write.table(c2glsdderesults,file="c:\\shenshaw\\Thesis\\Papers\\Paper3\
\gls\\c2glsdde.csv",sep=",",col.names=T,row.names=F)
write.table(c2glsdderesults1,file="c:\\shenshaw\\Thesis\\Papers\\Paper3
\\gls\\c2glsdde1.csv",sep=",",col.names=T,row.names=F)
```

Selected Statistical Code for Manuscript Three

Although complete analysis was performed on both sources of geocoding positional bias as well as the overall bias taking both these sources into account, only the code to analyze the overall bias is shown here.

```
# Reading in Data from text files

# Urban Offset
urban<-
read.csv("c:\\shenshaw\\Thesis\\Offset\\urbanNA.csv",header=T,sep=",")

# Rural Offset
rural<-
read.csv("c:\\shenshaw\\Thesis\\Offset\\ruralNA.csv",header=T,sep=",")

# Both urban and rural Offset
both<-
read.csv("c:\\shenshaw\\Thesis\\Offset\\bothNA.csv",header=T,sep=",")

# take out points where GPSed wrong

urban$GPSXMDft[urban$OffsetID==3|urban$OffsetID==18|urban$OffsetID==42]
<-NA
urban$GPSYMDft[urban$OffsetID==3|urban$OffsetID==18|urban$OffsetID==42]
<-NA

both$GPSXMDft[both$OffsetID==3|both$OffsetID==18|both$OffsetID==42]<-NA
both$GPSYMDft[both$OffsetID==3|both$OffsetID==18|both$OffsetID==42]<-NA

ugpsX<-urban$GPSXMDft
ugpsY<-urban$GPSYMDft
ugeoX<-urban$GeoXMDft
ugeoY<-urban$GeoYMDft

rgpsX<-rural$GPSXMDft
rgpsY<-rural$GPSYMDft
rgeoX<-rural$GeoXMDft
rgeoY<-rural$GeoYMDft

bgpsX<-both$GPSXMDft
bgpsY<-both$GPSYMDft
bgeoX<-both$GeoXMDft
bgeoY<-both$GeoYMDft

boffsetid<-both$OffsetID
uoffsetid<-urban$OffsetID
roffsetid<-rural$OffsetID

# change distance from meters to feet
udist<-urban$OFFSET..M.*3.28084
rdist<-rural$OFFSET*3.28084
bdist<-both$Offset*3.28084
```

```
# first get distance between geocoded x,y, and gps'd x,y

uoffset<-(sqrt((ugpsX-ugeoX)^2+(ugpsY-ugeoY)^2))
roffset<-(sqrt((rgpsX-rgeoX)^2+(rgpsY-rgeoY)^2))
boffset<-(sqrt((bgpsX-bgeoX)^2+(bgpsY-bgeoY)^2))

# take out points where arcview geocoded in wrong area altogether

roffset[roffset>50000]<-NA
boffset[boffset>50000]<-NA

# now calculate the new offset taking into account the laser distance
and offset

unewoff<-(sqrt(udist^2+uoffset^2))
rnewoff<-(sqrt(rdist^2+roffset^2))
bnewoff<-(sqrt(bdist^2+boffset^2))

# do student t tests of the data to see if there is a difference in the
mean b/t urban and rural

t.test(udist,rdist)
t.test(uoffset,roffset)
t.test(unewoff,rnewoff)

# look at the data using the hist function and summary statistics (only
one example shown)

hist(bnewoff)

Mbnewoff<-median(bnewoff,na.rm=T)
Nbnewoff<-length(bnewoff[!is.na(bnewoff)])
SDbnewoff<-sqrt(var(bnewoff,na.rm=T))/sqrt(Nbnewoff)
CILbnewoff<-Mbnewoff-1.96*SDbnewoff
CIUbnewoff<-Mbnewoff+1.96*SDbnewoff
Mbnewoff
Nbnewoff
SDbnewoff
CILbnewoff
CIUbnewoff
summary(bnewoff)
bnewoffover<-length(bnewoff[!is.na(bnewoff)&bnewoff>328])
bnewoffover

# now do variogram analysis, first make geodata objects with the data
(only one example shown)

library("geoR")

boffsetdata<-
data.frame(bgeoX=bgeoX,bgeoY=bgeoY,bgpsX=bgpsX,bgpsY=bgpsY,
      bdist=bdist,boffset=boffset,bnewoff=bnewoff,boffsetid=boffsetid)

bnewoffgeo<-as.geodata(boffsetdata,data.col=7,covar.col=8)

# look at variograms (only one example shown)
```

```
plot(variog(bnewoffgeo,max.dist=150000))
title(" variogram for bnewoff ")
plot(variog4(bnewoffgeo,max.dist=150000))
title("multidirectional variogram for bnewoff ")

############# fit variograms to models (one example ###################

plot(variog(bnewoffgeo))
plot(variog(bnewoffgeo,max.dist=100000,lambda=0))
lines.variomodel(cov.model="exp",cov.pars=
c(.7,10000),nugget=1.6,max.dist=100000)
title("Estimating a Variogram model for bnewoff  data")
bnewoff.remla<-likfit(bnewoffgeo,ini=c(.7,10000),
nugget=1.6,cov.model="exp",lambda=0,method="reml")
lines.variomodel(bnewoff.remla)
names(bnewoff.remla)
bnewoff.remla[1:3]

# conditional simulation with the new offset data (whites only)

locs.study<-ddegeo$coords
s.out<-output.control(n.predictive=1000)
bnewoffcs1<-krige.conv(bnewoffgeo,loc=locs.study,krige=krige.control
(obj.model=bnewoff.remla,lambda=0),output=s.out)
bnewoffcs<-bnewoffcs1$simulations

#Conditional simulation results into models from Manuscript Two to test
sensitivity. Only DDE is shown, but all models from Manuscript Two were
included in the loop.

# need to take already made geodata objects, and replace x, ys, and
distance to sites with new estimated residential locations

u<-complete.cases(c1data3[,c("dde","addX","addY")])
newc1data<-c1data3[u,]

n1<-length(newc1data$addX)
newx<-rep(NA,n1)
newy<-rep(NA,n1)
newdtosite<-rep(NA,n1)
newc1data<-
data.frame(newc1data,newx=newx,newy=newy,newdtosite=newdtosite)

sitex<-1108776
sitey<-725732

betas2<-numeric(0)
stder2<-numeric(0)
tstat2<-numeric(0)
rsqrds2<-numeric(0)
adjrsqrds2<-numeric(0)

for (i in 1:1000){
  r<-bnewoffcs[,i]
  angle<-runif(n1,min=0,max=2*pi)
  radius<- r
  dx<-cos(angle)*radius
```

```
dy<-sin(angle)*radius
newc1data$newx<-newc1data$addX+dx
newc1data$newy<-newc1data$addY+dy
newc1data$newdtosite<-sqrt((newc1data$newx-sitex)^2+
                            (newc1data$newy-sitey)^2)/5280

 u<-complete.cases(newc1data[,c("dde","newx","newy","newdtosite",
"age","gender2","water2","smoke","edu")])
tdde<-newc1data[u,]

covdde<-model.matrix(~newdtosite+age+gender2+edu+water2+smoke,
data=tdde)
covdde<-data.frame(covdde)
covdde<-covdde[,-1]
t2dde<-data.frame(tdde[,c(3:4,5)],covdde)
newgeodde<-as.geodata(t2dde,coords.col=1:2,data.col=3,covar.col=4:11)

newtrenddde<-formula(~newgeodde$covariate$newdtosite+newgeodde$
covariate$age+newgeodde$covariate$gender2+newgeodde$covariate$edu
+newgeodde$covariate$water21+newgeodde$covariate$water22+newgeodde$
covariate$water23+newgeodde$covariate$smoke)

locsdde<-newgeodde$coords

n<-dim(tdde)[1]
cov1<-c(220,30000)
nugget1<-180

newfitdde<-likfit(newgeodde, cov.model="exp", ini.cov.pars=cov1,
method="reml",nugget=nugget1, trend=newtrenddde)

newdmatdde<-as.matrix(dist(t2dde[,1:2]))
Cov<-cov.spatial(newdmatdde,cov.model="exp",cov.pars=
newfitdde$cov.pars)+diag(newfitdde$nugget,n)
Cov<-Cov/fic2t2btpcbspeaklip2$cov.pars[1]
Cinv<-solve(Cov)
X<-as.matrix(cbind(1,t2dde[,4:11]))
p<-dim(X)[2]
Y<-matrix(t2dde[,3],ncol=1)

beta.gls<-solve(t(X)%*%Cinv%*%X)%*%t(X)%*%Cinv%*%Y
var.gls<-diag(solve(t(X)%*%Cinv%*%X))
sse.gls<-t((Y-X%*%beta.gls))%*%(Y-X%*%beta.gls)
mse.gls<-sse.gls/(n-p)
gls<-cbind(beta.gls,sqrt(mse.gls*var.gls))

ssto<-t(Y)%*%Y-(1/n)*t(Y)%*%matrix(1,n,n)%*%Y
ddeR2.gls<-1-sse.gls/ssto
ddeR2adj.gls<-1-((n-1)/(n-p))*sse.gls/ssto

ddee.gls<-Y-X%*%beta.gls

betas2<-cbind(betas2,newfitdde$beta)
stder2<-cbind(stder2,sqrt(diag(newfitdde$beta.var)))
tstat2<-cbind(tstat2,newfitdde$beta/(sqrt(diag(newfitdde$beta.var))))
rsqrds2<-c(rsqrds2,ddeR2.gls)
adjrsqrds2<-c(adjrsqrds2,ddeR2adj.gls)
```

```
}

results2a<-betas2
results2b<-stder2
results2c<-tstat2
results2d<-cbind(rsqrds2,adjrsqrds2)

# now read results back in and look at distributions compared to
observed values found in Manuscript Two:

#read the files back in and get mean betas, pvalues, and adj rsquars #

ddebeta<-
read.csv("c:\\shenshaw\\Thesis\\Papers\\Paper2\\Results\\DraftResults\\
ddebetatog.csv",header=F,sep=",")
ddestder<-
read.csv("c:\\shenshaw\\Thesis\\Papers\\Paper2\\Results\\DraftResults\\
ddestdertog.csv",header=F,sep=",")
ddetstat<-
read.csv("c:\\shenshaw\\Thesis\\Papers\\Paper2\\Results\\DraftResults\\
ddetstattog.csv",header=F,sep=",")
ddersqrd<-
read.csv("c:\\shenshaw\\Thesis\\Papers\\Paper2\\Results\\DraftResults\\
ddersqrdtog.csv",header=T,sep=",")

ddebeta<-t(ddebeta)
dim(ddebeta)
row.names(ddebeta)<-NULL
ddebeta<-data.frame(ddebeta)
names(ddebeta)<-
c("Int","dtosite","age","gender","edu","water21","water22","water23","s
moke")
names(ddebeta)

ddestder<-t(ddestder)
dim(ddestder)
row.names(ddestder)<-NULL
ddestder<-data.frame(ddestder)
names(ddestder)<-
c("Int","dtosite","age","gender","edu","water21","water22","water23","s
moke")
names(ddestder)

ddetstat<-t(ddetstat)
dim(ddetstat)
row.names(ddetstat)<-NULL
ddetstat<-data.frame(ddetstat)
names(ddetstat)<-
c("Int","dtosite","age","gender","edu","water21","water22","water23","s
moke")
names(ddetstat)

# Here, the results for the table and figures in Manuscript Three for
the dieldrin model that are shown as an example.

# look at dieldrinlip
mean(dieldrinlipbeta$dtosite)
```

```
hist(dieldrinlipbeta$dtosite,xlab="Beta Coefficients of Distance to
Site Variable",main = paste(""))
abline(v=-1.586653)

t1<-hist(dieldrinlipbeta$dtosite,nclass=30)
lines(smooth.spline(t1$mids,t1$counts))
range(t1$counts)  # need for ylim
range(dieldrinlipbeta$dtosite)  # need for xlim
plot(0,0,ylim=c(0,17),xlim=c(-1.73,-1.46),pch=" ",
    xlab="Beta Coefficients of new Distance to Site
Variable",ylab="Frequency")
lines(smooth.spline(t1$mids,t1$counts))
abline(v=-1.586653)

title("Distribution of Beta Coefficients for New Distance to Site
Variable and Observed Value")
mean(dieldrinlipbeta$age)
hist(dieldrinlipbeta$age)
abline(v=-0.189)
mean(dieldrinlipbeta$gender)
hist(dieldrinlipbeta$gender)
abline(v=14.588)
mean(dieldrinlipbeta$edu)
hist(dieldrinlipbeta$edu)
abline(v=-1.772)
mean(dieldrinlipbeta$water21)
hist(dieldrinlipbeta$water21)
abline(v=221.17)
mean(dieldrinlipbeta$water22)
hist(dieldrinlipbeta$water22)
abline(v=9.39)
mean(dieldrinlipbeta$smoke)
hist(dieldrinlipbeta$smoke)
abline(v=21.796)

mean(dieldrinliptstat$dtosite)
hist(dieldrinliptstat$dtosite,xlab="t statistics of Distance to Site
Variable",main = paste(""))
abline(v=-2.048506)

t1<-hist(dieldrinliptstat$dtosite,nclass=30)
lines(smooth.spline(t1$mids,t1$counts))
range(t1$counts)  # need for ylim
range(dieldrinliptstat$dtosite)          # need for xlim
plot(0,0,ylim=c(0,15),xlim=c(-2.29,-1.89),pch=" ",
    xlab="t statistics of new Distance to Site
Variable",ylab="Frequency")
lines(smooth.spline(t1$mids,t1$counts))
abline(v=-2.048506)

mean(dieldrinliptstat$age)
hist(dieldrinliptstat$age)
abline(v=-0.67)
mean(dieldrinliptstat$gender)
hist(dieldrinliptstat$gender)
abline(v=1.67)
mean(dieldrinliptstat$edu)
```

```
hist(dieldrinliptstat$edu)
abline(v=-1.462)
mean(dieldrinliptstat$water21)
hist(dieldrinliptstat$water21)
abline(v=6.326)
mean(dieldrinliptstat$water22)
hist(dieldrinliptstat$water22)
abline(v=0.855)
mean(dieldrinliptstat$smoke)
hist(dieldrinliptstat$smoke)
abline(v=2.898)

mean(dieldrinliprsqrd$adjrsqrds8)
hist(dieldrinliprsqrd$adjrsqrds8)
abline(v=0.3756)

# Now get Variogram Figure of Error Data
variog(bnewoffgeo,max.dist=35000,lambda=0))
lines.variomodel(bnewoff.remla)
```

PERMISSION LETTER

JOHNS HOPKINS
BLOOMBERG
SCHOOL *of* PUBLIC HEALTH

615 N Wolfe Street • Baltimore, MD 21205 www.jhsph.edu

January 21, 2004

Ross Mullner, PhD
10301 S. Kostner
Oak Lawn, IL 60453

Dear Dr. Mullner:

I am completing a doctoral dissertation at Johns Hopkins University entitled "Spatial Attributes of Blood Organochlorine Concentrations in Washington County, Maryland: 1974-1989." Since the study entitled "Geostatistics and GIS: Tools for Characterizing Environmental Contamination" was part of my dissertation research, I would like your permission to reprint the following in my dissertation:

Henshaw, S. L., Curriero, F. C., Shields, T. M., Glass, G. E., Strickland, P. T., and Breysse, P. N. Geostatistics and GIS: Tools for Characterizing Environmental Contamination. Journal of Medical Systems (in press)

The requested permission extends to any future revisions and editions of my dissertation, including non-exclusive world rights in all languages, and to the prospective publications of my dissertation by UMI Company (formerly University Microfilm, Inc.). These rights will in no way restrict republication of the materials in any other form by you or others authorized by you. Your signing of this letter will also confirm that you own the copyright to the above-described material.

If these arrangements meet with your approval, please sign this letter where indicated below and return it to me in the enclosed return envelope. If you have any questions, please to not hesitate to contact me at: shenshaw@jhsph.edu or (410) 955-3094.
Thank you very much.

Sincerely,

Shannon L. Henshaw

PERMISSION GRANTED FOR THE USE REQUESTED ABOVE:

Ross Mullner   JAN 26, 2004

Ross Mullner , PhD   **Date**

Protecting Health, Saving Lives—*Millions at a Time*

CURRICULUM VITAE

Shannon L. Henshaw

Current Address

215 S. Washington St.
Baltimore, MD 21231
E-mail: shenshaw@jhsph.edu
Phone: 410-804-7413

Date and Place of Birth

April 17, 1977
St. Louis, Missouri
United States of America

Education:

- Bachelor of Science Degree in Environmental Geosciences: May 1999
  University of Notre Dame, School of Engineering;
  Department of Civil Engineering and Geological Sciences
- Master of Health Science Degree in Environmental Health Engineering and Industrial Hygiene: December 2001
  Johns Hopkins University, School of Hygiene and Public Health;
  Department of Environmental Health Sciences
- Current Doctor of Philosophy Degree candidate in Environmental Health Engineering
  Johns Hopkins University, Bloomberg School of Public Health;
  Department of Environmental Health Sciences

Work Experience:

- Environmental, Health and Safety Intern
  Unilever HPC-NA, Baltimore, MD 2000 – 2004
  - Responsibilities include a variety of safety, health and environmental tasks such as managing MSDSs, developing training presentations, writing manuals, helping with HazOps, conducting noise surveys, and conducting environmental and safety inspections.
  - Main focus has been leading a liquid waste minimization project in which I worked with several other employees to determine which manufacturing processes use the most water, waste the most water, and pollute the water by wasting product. The conclusions and recommendations from the study identified cost-effective improvements in raw material utilization, waste minimization and environmental impact.
  - Safety representative of the Ergonomics Team that consists of Union Safety Committee members, the facility nurse, and an engineer. Responsibilities of the team include ergonomically evaluating job tasks, and determining methods to improve the task to prevent worker injury.
- Teaching Assistant
  Johns Hopkins Bloomberg School of Public Health
  - Introduction to Environmental Health 2001, 2002, 2003
  - Principles of Industrial Hygiene 2002, 2003
- Environmental Consulting Intern
  Environmental Strategies Corporation, Pittsburgh, PA 1998
  - Responsibilities included regulatory research, data analysis, abatement technologies research, engineering calculations, interpreting geological maps, and entering data into a GIS system.
- Teaching Assistant
  University of Notre Dame School of Engineering 1999
- Laboratory Technician
  University of Notre Dame Rock Physics Laboratory 1997
- Personal Tutor 1994 – 1999

Special Training/Experience:

- American Industrial Hygiene Association Chesapeake Section Student Representative to the Executive Board, January 2001 to Present
- Environmental Health Sciences Student Organization (EHSSO) Divisional Representative to the Executive Board, March 2003 to Present
- Chair of the EHSSO Exam Committee, March 2003 to Present

- Completed the Johns Hopkins Risk Sciences and Public Policy Certificate, May 2001
- Johns Hopkins University Industrial Hygiene Student Association Vice President, September 2001 to September 2002
- Computer Skills in ArcGIS, R Statistical Computing Environment, word processing, Excel, Power Point, STATA, Matlab, Modflow, Maple, IslandDraw, and FORTRAN 90
- Participated in the Engineering Projects in Community Service (EPICS) Program through the University of Notre Dame servicing the City of Elkhart, Indiana as an environmental consultant, September 1998 to May 1999
- Participated in the Semester Around the World program through the University of Notre Dame, August to December 1996

  Studies concentrated in India

  Additional travel: Japan, Taiwan, Hong Kong, China, Singapore, Malaysia, Indonesia, Germany, Austria, Czech Republic, Italy, France, Spain, Switzerland, UK
- Participated in water relief and development project in Haiti through the University of Notre Dame, April 1999
- Six years of French education, September 1989 to June 1995

Honors:

Recipient of the 3M Industrial Hygiene Scholarship	2001
Recipient of the Stockhausen Occupational Health Scholarship	2001
Member of Tau Beta Pi Engineering Honors Society	1998 – 1999

Funding:

National Institute of Environmental Health Sciences (NIEHS) training grant	2001 – 2004
Johns Hopkins NIEHS Center for Urban Environmental Health grant	2003 – 2004
National Institute of Occupational Safety and Health training grant	1999 – 2000

Lectures and Presentations at the Johns Hopkins University:

- Doctoral Research Presentation — March 2004
  - "Spatial Attributes of Blood Organochlorine Levels in Washington County, Maryland, 1974-1989"
- Lecture in Spatial Statistics and GIS Course — March 2004
  - "Georeferencing and On-Screen Digitization in ArcGIS"
- Poster Presentation at the NIEHS Annual Workshop — February 2004
  - "Geostatistics and GIS: Tools for Characterizing Environmental Contamination
- Lecture in Noise and Other Physical Agents Course — November 2003
  - "Noise Control and Personal Protective Equipment"
- Proposed Doctoral Research Presentation — May 2002
  - "Spatial Attributes of Blood Organochlorine Levels in Washington County, Maryland, 1974-1989"
- Masters Essay Presentation — October 2000
  - "Liquid Waste Minimization in a Large Detergent Manufacturing Facility"

Publications:

Henshaw, S. L., Curriero, F. C., Shields, T. M., Glass, G. E., Strickland, P. T., and Breysse, P. N. Geostatistics and GIS: Tools for Characterizing Environmental Contamination. Journal of Medical Systems. (In Press.)

References/Curriculum Detail:

- Available upon request

JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH

615 N Wolfe Street • Baltimore, MD 21205 www.jhsph.edu

January 21, 2004

Ross Mullner, PhD
10301 S. Kostner
Oak Lawn, IL 60453

Dear Dr. Mullner:

I am completing a doctoral dissertation at Johns Hopkins University entitled "Spatial Attributes of Blood Organochlorine Concentrations in Washington County, Maryland: 1974-1989." Since the study entitled "Geostatistics and GIS: Tools for Characterizing Environmental Contamination" was part of my dissertation research, I would like your permission to reprint the following in my dissertation:

Henshaw, S. L., Curriero, F. C., Shields, T. M., Glass, G. E., Strickland, P. T., and Breysse, P. N. Geostatistics and GIS: Tools for Characterizing Environmental Contamination. Journal of Medical Systems (in press)

The requested permission extends to any future revisions and editions of my dissertation, including non-exclusive world rights in all languages, and to the prospective publications of my dissertation by UMI Company (formerly University Microfilm, Inc.). These rights will in no way restrict republication of the materials in any other form by you or others authorized by you. Your signing of this letter will also confirm that you own the copyright to the above-described material.

If these arrangements meet with your approval, please sign this letter where indicated below and return it to me in the enclosed return envelope. If you have any questions, please to not hesitate to contact me at: shenshaw@jhsph.edu or (410) 955-3094.
Thank you very much.

Sincerely,

Shannon L. Henshaw

PERMISSION GRANTED FOR THE USE REQUESTED ABOVE:

JAN 26, 2004

Ross Mullner , PhD **Date**