A Thesis

entitled

Planning for Spatial Analysis of Links between Parkinson Disease and Pesticide Exposure

by

Xuan Li

Submitted to the Graduate Faculty as partial fulfillment of the requirements for the

Master of Arts Degree in Geography

______________________________________________

Dr. Kevin Czajkowski, Committee Chair

______________________________________________

Dr. David J. Nemeth, Committee Member

______________________________________________

Dr. Bhuiyan Alam, Committee Member

______________________________________________

Dr. Sadik A Khuder, Committee Member

______________________________________________

Dr. Patricia Komuniecki, Dean

College of Graduate Studies

The University of Toledo

August 2011
An Abstract of
Planning for Spatial Analysis of Links between Parkinson Disease and Pesticide Exposure

By

Xuan Li

Submitted to the Graduate Faculty as partial fulfillment of the requirements for the
Master of Arts in Geography

The University of Toledo
August 2011

Parkinson disease is a debilitating disease that affects the central nervous system of patients. According to the report from The National Institute of Neurological Disorders and Stroke, the United States has over 500,000 people with Parkinson disease, and the number is increasing by 50,000 every year. Previous studies usually focused on risk factors of Parkinson disease, including genetic make-up and living habits. Recently, some researchers have conducted studies focusing on environmental impact issues with some reports indicating a high prevalence rate of Parkinson disease among farmers.

The purpose of this research is to develop a plan for Ohio to implement the Geographic Information System with spatial analysis to explore the association between potential environmental links with Parkinson disease. The relationship between Parkinson disease and pesticide applications in previous studies was reviewed. Four methods can be used in the study: Cross-sectional, Ecologic, Cohort and Case-control study. The results indicate an ecologic study with ZIP code unit would be the most appropriate type of study.
given the datasets we have. The function of Geocoding cannot be used in this study, because of missing patients’ addresses. However, spatial analysis can still be implemented by using the mended GIS-based model discussed by Nuckols et al. (2007). Future directions are also discussed at the end of this thesis.
Acknowledgements

I would like to take this opportunity to express my gratitude to my advisor, Dr. Kevin Czajkowski. His guidance, encouragement, patience, advice and financial support make the completion of this thesis possible. I also want to thank Dr. David J. Nemeth, Dr. Sadik A Khuder and Dr. Bhuiyan Alam for their advice and contribution to my thesis.

I would like to thank April Ames for her kindness, help and encouragement. Thanks to Xueying Chen, David Dean, and Elaine Reynolds. I appreciate your patience and help during the work. Thanks to Jie Ren, Xi Li, Han Li and Qifeng Wang. Thank you for your friendship and spirit support. I also want to thank all faculty and staff in our department.

Finally, I would express my love to my parents, my grandma, my husband and my son. I don’t know how to tell you my deepest sense of gratitude and appreciation. Thanks for your endless love.
Contents

Abstract ........................................................................................................................................ iii

Acknowledgements ...................................................................................................................... v

Contents ....................................................................................................................................... vi

List of Tables ............................................................................................................................... ix

List of Figures .............................................................................................................................. x

1 Introduction ................................................................................................................................. 1

1.1 Background ............................................................................................................................ 1
1.2 Problem Statement .................................................................................................................. 3
1.3 Objectives ............................................................................................................................... 4

2 Literature Review .................................................................................................................... 5

2.1 Background of Parkinson Disease ......................................................................................... 5
2.2 Epidemiologic Studies of PD .................................................................................................. 6
  2.2.1 Case Reports and Case Series ........................................................................................ 6
  2.2.2 Cross-sectional Studies ................................................................................................... 7
  2.2.3 Cohort Studies ................................................................................................................ 9
  2.2.4 Case-control Studies .................................................................................................... 12
3 Geographic Information System (GIS) in Health Studies

3.1 Geographic Information System

3.2 Pesticide Exposure Assessment
   3.2.1 Geocoding Method
   3.2.2 Data for Assessment
   3.2.3 Ambient Pesticide Exposure Estimates

3.3 Statistical Software
   3.3.1 R Software
   3.3.2 SAS and SPSS Software

3.4 Statistical Methods

4 Methodology

4.1 Study Area
4.2 Dataset
4.3 Data Sources
4.4 Data Analysis
4.5 Research Methodology

5 Results

5.1 Comparison of Cross-sectional, Ecologic, Cohort and Case-control Studies
5.2 Comparison of Geocoding Methods
5.3 Comparison Methods of Pesticide Exposure Estimates
5.4 Comparison of Statistics Software
5.5 Data Analysis
   5.5.1 Study Design and Study Unit Selection
   5.5.2 Variables Selection and Calculation
   5.5.3 Statistical Analysis Procedure

6 Conclusion and Discussion

6.1 Conclusion
6.2 Future Study
List of Tables

2.1 Sample of Method to Calculate the Odds Ratio 13
2.2 Demographic Characteristics of the Study Population 16
4.1 General Information about Patients’ Age in the Dataset 42

5.1 Advantages and Disadvantages of Cross-sectional, Ecologic, Cohort and Case-contrStudies 49
5.2 Advantage and Disadvantage of Geocode Methods 50
5.3 Advantage and Disadvantage of Geocoding Individual Data Methods of McElroy et al. (2003) and Gilboa et al. (2006) 51
5.4 Comparison Three Pesticide Exposure Estimate Methods 52
5.5 Advantages and Disadvantages of R, SAS and SPSS Software 53
List of Figures

Figure 2-1: Sample Each Step of Subject Selection ..................................................8

Figure 2-2: Example of Every Step of Subjects’ Data Collection ...............................10

Figure 2-3: Example of Level of Exposure Referred to Specific Job Titles in Baldi et al..............................................................11

Figure 3-1: Example of Dot Density Map .................................................................19

Figure 3-2: Count of births by Census Block Group and ZIP code .........................20

Figure 3-3: Example of “Density equalized map” ......................................................21

Figure 3-4: Example of Overlay Using the Map of Springfield, Massachusetts ........22

Figure 3-5: Sample of the reclassification function served in the aim of Prevention by Reclassifying the Same Data. (The first map shows the DWI .................23

Figure 3-6: Example Geocoding Procedure Developed by McElroy et al. .................26

Figure 3-7: Flow Chart of Bell et al (2001) Pesticide Exposure Estimate Method ......31

Figure 3-8: Flow chart of Rull et al (2003) Pesticide Exposure Estimate Method.......32

Figure 3-9: Flow Chart of Nuckols et al (2007) Pesticide Exposure Estimate Method ...33

Figure 3-10: Example of Step 5 of Nuckols’ study .....................................................34

Figure 4-1: Study Area: Ohio Base Map .................................................................39

Figure 4-2: Land Cover of Ohio State .................................................................40
Figure 4-3: Box Plot of Age in the Data Set ..............................................42

Figure 4-4: Example of Private Pesticide Applicator Restricted Use Recordkeeping .....44

Figure 5-1: Locations of the 8 Counties of the Dataset ....................................54

Figure 5-2: Distribution of PD Patients’ Density Map in Montgomery County by ZIP code ........................................................................................................55

Figure 5-3: Distribution of PD Patients’ Density Map in Franklin County by ZIP code ........................................................................................................55

Figure 5-4: Distribution of PD Patients’ Density Map in Hamilton County by ZIP code ........................................................................................................56

Figure 5-5: Distribution of PD Patients’ Density Map in Cuyahoga County, Portage County, and Mahoning County by ZIP code ................................56

Figure 5-6: Distribution of PD Patients’ Density Map in Lucas County and Wood County by ZIP code .................................................................57

Figure 5-7: Flow chart of pesticide exposure assessment method ......................59
List of Abbreviations

CDWR.................................................. California Department of Water Resources

GHC .................................................... Group Health Cooperative

GIS...................................................... Geographic Information System

IPD...................................................... Idiopathic Parkinson’s Disease

OC...................................................... Organochlorine

OR...................................................... Odds Ratio

PD...................................................... Parkinson Disease

PLS...................................................... Public Land Use

PLSS.................................................... Public Land Survey System

PUR...................................................... California Pesticides Use Reports

SPSS.................................................... Statistical Package for the Social Sciences

TIGER.................................................. Topologically Integrated Geographic Encoding

and Referencing system
Chapter 1

Introduction

1.1 Background

“I done wrassled with an alligator. Tussled with a whale. I done handcuffed lightening, threw thunder in jail. I murdered a stone, I hospitalized a brick, I'm so mean I make medicine sick.”

- Muhammad Ali

The words above were said by Muhammad Ali-- history's most famous boxing athlete. He was the World Heavyweight Champion three times and was considered as one of the greatest heavyweight championship boxers of all time. Ali was well known for his unorthodox boxing style. "float like a butterfly, sting like a bee" is how he has described his fighting style. However, these outstanding characteristics began to fade gradually. His reflexes became slowed. His speech became diminished in strength and quality. But his mind was still clear and sharp. After several years his condition was getting worse and worse. Finally, Ali was diagnosed with Parkinson disease in 1984 (Hauser, 1991).
Ali was not the first person diagnosed with PD. People have known Parkinson disease as a condition since ancient times. It was referred by Kampavata in the ancient Indian medical system. Parkinson disease also had been described by a physician as “shaking palsy” in AD 175. In 1817 it was defined by Dr. James Parkinson for the first time.

Parkinson disease (PD) is a degenerative disorder of the central nervous system that often impairs the sufferer's motor skills, speech, and other functions (Jankovic, 2008). It impacts the central nervous system negatively and reduces the brain’s ability to co-ordinate movement. It is a complex movement disorder and the second most common neurodegenerative disorder affecting elderly people after Alzheimer’s disease (Beate et al. 2006).

The number of PD patients is increasing rapidly. The overall incidence of Parkinson's disease, based on several worldwide studies, is about 10-20 cases per 100,000 people per year (Blackmer, 2010). Prevalence estimates tend to vary, but the currently accepted figure is approximately 100-200 cases per 100,000 people. Furthermore, the frequency of Parkinson's disease in the United States is similar to that found internationally (Blackmer, 2010).
1.2 Problem Statement

With the growing aging population, the number of elderly people who are living with PD has increased. It is reported that it was increased to about 10-20 new PD patients per 100,000 people each year (Blackmer, 2010). PD can cause the quality of patients’ lives to decrease. Since PD patients need more nursing, it will cost more to take care of them. All these factors are causing PD to get more attention from epidemiological researchers. For more than two decades, reports have suggested that pesticides and herbicides may be an etiologic factor in idiopathic PD (Lockwood, 2000 and Priyadarshi et al., 2000). Ritz and Costello (2006) have conducted a study about the relationship between the agricultural environment and increasing risk of PD. A strong connection between PD and pesticides exposure was found (Ritz, and Costello, 2006). However, the research conducted to date was limited by small sample size. Their results could not strongly support their hypothesis. Hence, it would be responsible to conduct more studies on other areas to verify the hypothesis.

The primary research purpose of this study is to make a plan which is suitable for the spatial analytic study of exploring the relationship between Parkinson’s disease and pesticide exposure in Ohio using a Geographic Information System (GIS).
1.3 Objectives

The objectives of this study are as follows:

1) Determine how existing studies of Parkinson disease are conducted; what has been found by researchers; how GIS is used in environmentally linked illness studies; and what methods are used to conduct this research;

2) Compare, discuss and evaluate advanced methods to analyze Parkinson’s disease.

3) Develop a plan to implement GIS with spatial analysis in order to explore the association between potential environmental links with PD.
Chapter 2

Literature Review

2.1 Background of Parkinson Disease

Parkinson disease negatively affects the patients’ activities of daily living. The patients also may not be able to work when their condition gets worse. Recently, with the increasing population of elderly people in Western countries and the high cost of patient care, PD has received much attention in the public health research area (Brown, 2006).

Generally, most studies about PD focus on therapeutic medicine as well as developing treatment such as physical therapy, lifestyle modification etc. Ho et al. (1989) conducted a population-based case-control study. The researchers focused on exploring whether the occupational use of herbicides can be a risk factor of PD. The results showed that PD had occurred at a high rate among farmers. Furthermore, people who lived in or nearby rural areas appeared to have a higher risk of getting PD. This finding has contributed to a possible hypothesis that the application of pesticides in agricultural areas may be one of the causal factors of PD. Some researchers have changed their focus area to study the relationship between pesticide application and PD to verify that hypothesis.
A recent study conducted by Ritz and Costello (2006) used a GIS-based exposure assessment model to analyze the association between pesticide application and PD. Spatial analysis principles were also employed in their study. They concluded that the GIS – based model was valid for estimating residential exposures to agricultural pesticides.

2.2 Epidemiologic Studies of PD

Many epidemiologic studies investigated the link between pesticides exposure and PD (SemchukLove, and Lee, (1992); Brown, (2006); Ritz and Costello, (2006); Priyadarshi, Khuder, Shaub, and Shrivastava, (2000) etc. Brown et al. (2006) separated these studies into four categories: case reports, cross-sectional studies, cohort studies, and case-control studies.

2.2.1 Case Reports and Case Series

Generally, case report and case series studies use the clinical route to identify the disease (Kramer, and Boivin, 1988). Six cases which related transient and reversible Parkinsonism after acute exposure to pesticides have been reported by Muller-Vahl et al. (1999) and Bhatt et al. (1998). All of the patients showed similar symptoms to those who experienced Parkinson’s disease after attempted suicide (Brown, 2006).

Two other cases have been reported by Shahar et al. (2001) and Arima et al. (2003). The first case reported by Shahar et al. (2001) presented a 17-year-old female who developed acute extra-pyramidal Parkinsonism after attempting to commit suicide. The second case reported by Arima et al. (2003) described an 81-year-old woman who also had
consumed an insecticide with a suicide attempt. She was admitted for treatment relating to acute severe organophosphate poisoning (Arima et al. 2003).

These reported cases implied that organophosphate poisoning could lead to symptoms similar to those experienced by PD patients. Also, the neurotoxicity associated with organophosphate poisoning could have an extreme impact on the central nervous system.

Rajput et al. (1986 and 1987) evaluated twenty-one PD patients’ childhood environment, whose symptoms began before 40 years. All of the PD patients were born and raised in Saskatchewan; nineteen lived in rural Saskatchewan for their first 15 years. Rajput et al. (1986) notes that “detailed population analysis indicates a strong predisposition to early onset idiopathic Parkinson’s disease (IPD).” Rajput et al. (1986) also concluded that “rural Saskatchewan environments contribute to IPD and that well water used in childhood should be considered as a potential vehicle for the etiological agent.”

2.2.2 Cross-sectional Studies

A cross-sectional study is also called a prevalence study or health survey, which can describe the frequency of the disease at a specific point in time. Barbeau et al. (1987) investigated the distribution relating to the prevalence of PD in 9 rural regions by using “the unique combination of a homogeneous genetic and racial origin in the rural population of Quebec and the facilities of free and universal access to medical care”. They found an
association between Parkinson’s disease and the geographical distribution of pesticide application. Engel et al. (2001) indicted that the general use of pesticides also could increase the risk of Parkinson disease, but the authors had not tried to clarify which category of pesticides could increase the risk or if all pesticides can raise the risk of Parkinsonism. Engel et al. (2001) obtained subjects’ information from a project conducted by the Washington State Department of Health from 1972 through 1976. Then, they used the information to contact all 739 subjects from that project and 185 participated (Figure 2-1).

![Figure 2-1: Sample Each Step of Subject Selection (Engel et al. 2001)](image)

Engel et al. (2001) collected information about pesticide use ascertainment by using a self– administered questionnaire and gathered results based on data analysis.

Researchers conducted more studies (Bennett et al. 1988 and Sala et al. 1999) on pesticide exposure. The results showed no links between pesticide exposure and PD. However, several studies (Granieri et al. 1991; Lee et al. 2002; Schulte et al. 1996; Wang et al.1994) still found that the occupation of farming and rural living could be risk factors of
PD.

2.2.3 Cohort Studies

A cohort study observes two groups (One group’s patients have a certain condition and the other group’s patients have not been affected by that certain condition) over a period of time and compares the outcomes (Samet, J M. and Muñoz A. 1998). The purpose of a cohort study is to explore associations and temporal relationships between hypothesized risk factors and the specific disease. The first step of a cohort study is to record healthy subjects, who are under or not under the exposure environment of the putative risk factors. The subjects with exposure to hypothesized risk factors are followed during the study. Their health status are also watched and recorded accurately on a schedule. In order to ensure the accuracy of the and improve the comparison, the subjects, who do not live in the exposure environment, are studied during the same time as the subjects with the exposure.

The incidence rate is “the proportion of subjects who develops the disease under study within a specified time period” (Meirik 2008). The association can be measured by the relative risk, which is the incidence rate of the subjects with exposure divided by the incidence rate of subjects without exposure (Meirik 2008). When the relative risk is equal to 1.0, the incidence rate is the same among exposure and non exposure groups. It also indicates the exposure is not a risk factor of that disease. When the relative risk is more than 1.0, it means that exposed people have a higher risk of getting the disease than
non-exposed subjects. Conversely, when the relative risk is below 1.0, it shows the exposure could have a protective factor.

Baldi et al. (2003) conducted a cohort study which indicated that the occupation associated with farming or agriculture (which was believed a proxy for exposure to pesticide) could be an important causal factor of PD. This cohort study involved 1507 elderly people, whose age was over 65 years old (Figure 2-2).

![Figure 2-2: Example of Every Step of Subjects’ Data Collection (Baldi et al. 2003)](image)

Detailed occupational historical information of all subjects was collected through a face-to-face interview. All subjects have completed structured questionnaires during the interview. Baldi et al. (2003) divided the exposure into occupational exposure and environmental exposure. The occupational exposure to pesticides was based on job title and assessed by “a panel of six experts blinded to neurologic” (Baldi et al. 2003). A level of exposure matrix was developed at the end of assessment by different job titles. (Figure 2-3).
Researchers hypothesized environmental exposure based on the subjects’ residential places: rural residency and residency in a district with vineyards. Then, they analyzed the effect of occupational and environmental pesticides exposure on men and women separately. The results showed that the rates of Standardized hospitalization and incidence were both increased among agricultural workers (Baldi et al. 2003).

Another cohort study was conducted by Petrovitch et al. (2002) on the island of Oahu, Hawaii. It was a follow-up study that lasted 30 years. In addition, 7986 Japanese
American men, who were born between 1900 and 1919, were selected as the subjects in this study from the Honolulu Heart Program. They collected pesticides exposure information of all subjects by completing a questionnaire. Then, they estimated the crude and age-adjusted incidence rates of PD based on the years worked on plantations. The result showed that the higher number of years of plantation work led to higher risk of developing PD.

2.2.4 Case-control Studies

A case-control study is a study which compares the proportion of certain exposure or particular characteristic of interest of two groups (Armenian, 1994). The subjects of one group who have a specific disease are called “case group”; by comparison, the people in the other group without that disease are called “control group”.

Compare to cohort studies, due to a lack of data, the incidence rate cannot be calculated in case-control studies and neither the relative risk. However, the odds ratio, which is “the ratio of odds of exposure in diseased subjects to the odds of exposure in the non-diseased” (Meirik, 2008), can be calculated for measuring the association between exposure and disease in the case-control studies. Table 2-1 (Meirik, 2008) shows the method of calculating the odds ratio in a case-control study. The meaning of the value of the odds ratio is the same as the relative risk. If the odds ratio is equal to 1.0, it means the exposure cannot be a risk factor of that disease; If the odds ratio is more than 1.0, it means that people with exposure have a higher risk of getting the disease than non-exposed
subjects; Conversely, if the odds ratio is below 1.0, it means the exposure is a protective factor for the disease (Meirik, 2008).

Table 2-1: Sample of method to calculate the odds ratio (Meirik, 2008).

<table>
<thead>
<tr>
<th>Exposure</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes(cases)</td>
</tr>
<tr>
<td>Yes</td>
<td>a</td>
</tr>
<tr>
<td>No</td>
<td>c</td>
</tr>
<tr>
<td>Odds of exposure</td>
<td>a/c</td>
</tr>
<tr>
<td>Odds ratio</td>
<td>a/c:b/d</td>
</tr>
</tbody>
</table>

Ho et al. (1989) conducted a case-control study in Hong Kong. The researchers tried to explore the association between environmental factors related to agriculture activities and PD. Two districts of Hong Kong were selected as the research area. In these two districts 3.4% of people have Parkinson’s disease. First, Ho et al. (1989) conducted a survey of 561 elderly residents. People who participated in this study were separated into two groups - case group and control group. The “case group” contains the people with Parkinson’s disease. The “control group” contains the elderly people without PD. They interviewed all subjects by a face-to-face interview following a structured questionnaire. The aim of the interview was to collect the information of subjects’ social characteristics, medical history, health practices, and environmental factors, including rural living, farming activities, previous exposure to pesticides and herbicides and consumption of raw
vegetables. They used odds ratios to “estimate the relative risk of the various studied risk factors”. The chi-square test for testing the trend of dose-response relationships was also used in this study. Finally, Ho et al. (1989) concluded “subjects with residence of long duration in rural areas, with engagement in farming, with previous use of herbicides and pesticides, and with habitual consumption of raw vegetables had a statistically significantly increased risk of Parkinson’s disease.”

Another case-control study, with the same objective as Ho et al (1989), was conducted by Koller et al. (1990) in Hong Kong. This study involved 150 patients who were randomly selected from the Movement Disorder Clinic. Another 150 people were found from neurologic and medical clinics, whose age and sex are similar with the patients group, to make the control group for comparison. Koller et al. (1990) implemented a survey by face-to-face interviews using a questionnaire. The information collected included living histories, sources of drinking water, and occupations. After that, they collected pesticide exposure data, which including the type of the exposure, how many years of exposure, how much was applied to each acre, and which chemicals were used on which type of crop. Finally, they performed the statistical analysis. The “findings” failed to identify relationships between pesticide and Parkinson’s disease.

A population-based case-control study has been conducted by Semchuk et al. (1992) in Calgary, Canada. The researchers focused on the risk factors of the agricultural related activities and pesticide exposure. The study involved 130 patients of residents of Calgary
as the cases and 260 people without PD were randomly selected as controls with age-and sex-match. Semchuk et al. (1992) narrowed down the questions to ask more details about the occupational histories, for instance, they recorded “the start and stop dates and details about outdoor work and any chemicals or compounds used on the job, including the dates on which the chemicals were first and last used.” The findings of the study indicated that the people who had an agricultural work experience or had to use pesticide chemicals during work had more risk to become a patient with PD.

Firestone et al. (2005) performed a population–based case–control study to explore associations between pesticide exposure and PD. Two hundred and fifty PD patients were selected from Group Health Cooperative (GHC) in Washington State, who identified the PD patients through 1992 to 2002. All control subjects, 388 healthy people, were randomly selected from GHC and did not have historical exposure, occupational exposure or residential locus. Both case and control groups have similar frequency of age, sex, GHC location, and original year of GHC enrollment. Table 2-2 describes the demographic characteristics of the subjects in this study.
Information about demographic, occupation history, medical, pesticide exposure, residential history, and so on were obtained using a structured interview. The results showed that home-based pesticide application has no association with PD. However, Firestone et al. (2005) found drinking well water could be a causal factor of PD.

Ecologic study is also an observational study. An ecologic study is a study which uses groups as the study unit to compare and analyze the variables at aggregated level. The scale of study unit is various. It can be a school, a community, a ZIP code area, a county and even a country. The data used in an ecologic study is usually already collected and ready for use. Therefore, the research can be implemented in a short time and also save much more money than the other studies. However, the relationship observed between variables on a group level cannot represent the association at an individual level that may lead to bias.
Chapter 3

Geographic Information System (GIS) in Health Studies

3.1 Geographic Information System

GIS has become a powerful tool in many different fields, which includes geography, geology, environmental studies, business, and criminal justice (McCall, 2003). Researchers can get visual information about the distribution of certain diseases in a new way. They can also inquire into some relationships, which are not previously seen with traditional epidemiological method, by using the GIS technology in the study (McCall, 2003). Due to its visual power, GIS maps become symbols for the social and environmental conditions that are “contained” in a geographic square (Cromley and McLafferty, 2002). It can also provide “a digital lens for exploring the dynamic connections between people, their health and well-being, and changing physical and social environment” (Cromley and McLafferty, 2002). In addition, the functions of integrating and analyzing spatial data make GIS an excellent analytic tool for public health studies.

Epidemiologists may focus on certain questions, such as “where do patients live?”,
“where are the agents of disease?”, “how is the disease distributed?” etc. All the questions above are related to analyzing the disease in terms of spatial factor. Spatial epidemiology is becoming a new discipline which associates with analyzing geographical variations of the disease, particularly with regard to variations in environmental exposures at the small-area scale (Rushton, 2003). The powerful functions of the Geographic Information System (GIS) can solve these research questions and make the study of illness visible.

The traditional studies of epidemiology often use demographic or socioeconomic information to calculate specific category and standardized incidence rates. A Regression model was generally used to examine the hypothesized risk factors at the time, but these studies failed to analyze spatial distribution of the disease or use visualization techniques. The four categories of GIS specific functions can supplement the lack of spatial analysis in epidemiological studies.

Bailey and Gattrell (1995) illustrated the four categories of GIS specific functions as data integration and management, visualization, exploratory spatial analysis, and mathematical modeling. These functions play a significant role in spatial epidemiological studies.

Data Integration is a function of GIS for integrating data from many existing sources decreasing the need for primary data collection for new studies (McCall, 2003). McCall (2003) described that GIS also had mathematical functions to generate the integrated data to create new variables or estimate values for existing variables.
The second category of visualization is mapping. With the process of geocoding, dot density maps of specific disease cases, such as cancer studies by exact location, can be automatically generated (McCall, 2003). By using geocoded data, GIS can count the total numbers of a geographic area and group the underlying population of that area to determine prevalence.

Figure 3-1 is an example of dot density map made by GIS. It shows the H3N2 swine epidemic situation in Ontario in 2005. Every dot in the map represents 7 herds in the target population. This map clearly shows how H3N2 swine epidemic is distributed in the study area.

Figure 3-1: Example of Dot Density Map
Figure 3-2 shows the number of births in Duval County by census block and ZIP code area. It used the graduate color in GIS to classify the total number of births from a small number to a large number.

![Figure 3-2: Count of births by Census Block Group and ZIP code](http://mch.peds.ufl.edu/gis_maps/duval.html)

Another advantage of GIS is smoothing techniques, which is used to get rid of the irregularities seen in 2D mapping and can be especially applied in mapping cancer incidence rates. Figure 3-3 “Density equalized map” is created through kernel smoothing by Selvin et al. (1998), which is used to describe late-stage breast cancer incidence on a continuous three-dimensional surface without regional boundaries.
Besides smoothing techniques, the functions of point and polygon overlay, and buffering can provide visualization to show the relationship between environmental exposure and diseases in GIS. For example, Figure 3-4 shows overlay of cases layered with Hydrography. From this map, the relationship between cases and hydrography is very obvious; cases are distributed along the hydrography area.
The third function category, exploratory spatial analysis, is based on the results of visualization and examines whether visualized patterns or relationships occur by chance or not. Since the exact locations of cases can be mapped, Investigators primarily apply exploratory spatial analysis on the testing for clustering of cancer cases (McCall, 2003). The last category of mathematical modeling can be used to estimate the relationship between various factors, or to predict unknown values.

Lipton et al. (2003) performed a study about exploring the relationship between alcohol stores’ locations and alcohol-related problems, especially violence, to better exemplify the application of spatial analysis by using GIS. First, they introduced some context of spatial analysis focused on prevention issues in public health. They described it
as “a potentially powerful approach” which demanded very complex data and implementation. Then, to better investigate the geographic relationship between the density of liquor stores and violent crime, they analyzed the general context of the distribution of liquor stores (Lipton et al. 2003). They also described the methods which were used in previous research for this issue. Finally, Lipton et al. (2003) used a specific example to show how to study the alcohol problem issues on the community scale. The essential key functions of GIS: geocoding, data overlays, reclassification (Figure 3-5), and distance/adjacency measures were also applied in Lipton et al. (2003) study.

Figure 3-5: Sample of the reclassification function served in the aim of Prevention by Reclassifying the Same Data. (The first map shows the DWI Conviction Rate by four different categories. The second map demonstrates this rate as a single category, but only shows the highest 10% rate areas)
The spatial analyst function of GIS helps researchers to conduct studies related to geographic distribution. It is becoming a popular approach to use GIS spatial analysis functions to study Parkinson’s disease associated with the risk factor of pesticides exposure. Recently Ritz and Costello (2006) conducted a study in California. This study not only focused on the relationship between pesticide exposure and Parkinson’s disease, but also employed a Geographic Information Systems-based pesticides exposure assessment model.

Ritz and Costello (2006) developed the GIS – based methods in their study to assess exposure of pesticides. The researchers geocoded all subjects’ residential histories’ addresses. The information of subjects’ residential histories was collected using a questionnaire, which include start-end dates of residence and exact street addresses. Ritz and Costello (2006) also created “a cumulative exposure score for a given residence based on the weighted average of organochlorine (OC) applications in a Public Land Use (PLS)”.

There were some limitations in the study. The statistical accuracy was restricted by small sample size. In addition, Ritz and Costello (2006) didn’t consider wind patterns and the equipment type of pesticide application. However, the results showed the specificity of the model was good, which could reduce misclassification bias of pesticide exposure in case-control settings commonly applied to study rare diseases (Ritz and Costello. 2006).
3.2 Pesticide Exposure Assessment

Pesticide exposure assessment is an essential part of studying the links between pesticide exposure and PD. Generally, the spatial analysis method for assessing pesticide exposure includes three steps: geocoding all subjects’ addresses, preparing data for assessment, and ambient pesticide exposure estimates.

3.2.1 Geocoding Method

Some functions of GIS, such as geocoding, data overlays, buffering, reclassification, and distance/adjacency measures, are very useful for inquiring into the relation between geographic location and certain disease (McElroy et al. 2003; Lipton et al. 2003). The utility of GIS relies on completeness and accuracy of geocoding (McElroy et al. 2003).

Geocoding also called address matching is the procedure of delegating research subjects’ residences latitude/longitude coordinates that closely approximate their true locations (Gilboa et al. 2006; McElroy et al. 2003). Yang et al. (2004) described that “geocoding provides a basis for addressing questions applicable to many areas of interest – where a certain population resides, where needs are located, where resources exist, and where to target resources?”

The scientists put significant effort into geocoding accuracy, which can impact the outcomes directly. An incomplete geocoding can weaken the power of statistic analysis. Even more, it can lead to an incorrect result. Generally, in pesticide exposure assessment
procedure, geocoding means geocode subjects’ or cases’ addresses in epidemiological study. Since the distance between cases and estimated exposure area can be a significant variable for exploring the hypothesis association and geocoding is used for locating cases’ addresses, geocoding is very important for this study. It can directly impact the accuracy of final results.

A significant factor of geocoding is the matching rate of participants’ addresses. There are many problems that can cause low match rates in geocoding, such as a variation in spelling an address from US Postal Service, a changed street number, or an incomplete street reference map (McElroy et al. 2003). McElroy et al. (2003) developed an iterative method of geocoding (Figure 3-6) and improved the matching rate up to 97% in a large population-based study.

![Figure 3-6. Example Geocoding Procedure Developed by McElroy et al. (2003)](image-url)
Figure 3-6 shows each step of geocoding conducted by McElroy et al. (2003). The researchers finished all procedures using eight steps. The first step is address improvement. They changed the addresses’ format to correspond to the US Postal Service format. For example, “2627 North 86 street” was changed to “2627 N 73rd St.” and “Campus village apt 11, 1702 Secor Road” was changed to “1702 Secor Rd.”. After step 1, 12,950 addresses were successfully changed. These 12,950 changed addresses were directly used to match geocoding, using 2000 and 1995 Topologically Integrated Geographic Encoding and Referencing (TIGER) street maps. These 2000 and 1995 TIGER street maps were used as the reference street map files in step 2. The other 1,854 participants’ addresses, which were equivocal and needed more information to identify, were put into step 3. After step 2, there are still 3796 (1915 from step 2 and 1854 from step 1) that remained unmatched. Compare to total 14,804 mailing addresses, these 3796 unmatched addresses cannot be ignored. To solve this problem, Internet mapping engines were used to improve the geocoding matching rate (step 3). For instance, a participants’ information (eg, telephone number and name) can be entered into an Internet mapping engine such as Anywho.com, to update the street addresses. And then, the researchers re-contacted the subjects in order to get accurate address information (step 4). After these four steps, only 470 addresses were not matched (3%). Finally, these 470 unmatched addresses were geocoded to the zip code centroid. For the multiple matched addresses, hierarchical rules were employed to designate the single latitude/longitude coordinate per participant.
In McElroy et al. (2006) study, by using Internet mapping engines and successfully re-contact subjects, the researchers increased the matching rate to 97% and solved the problem of urban bias which is misgeocoding of the participants from rural areas. In addition, compare to re-contact each unmatched subject, using Internet map engines is less expensive and timesaving. However, there are still unavoidable factors which can impact the accuracy of geocoding. For example, since the participants may move without updating their addresses, the addresses from the Internet may be inaccurate. Also, the street map they used may contain some degree of positional inaccuracy which can reduce the matching rate.

Another population-based case-control study (Gilboa et al. 2006), which involved investigating the relationship between exploring air pollution exposure during the third to the eighth weeks of pregnancy and the risk of selected birth defect, used a different method to improve the matching rate of geocoding. The addresses with unsure information were excluded before geocoding. The 5338 cases and 4574 control were finally selected. Comparing the method used in McElroy et al. (2003), the method in Gilboa et al. did not employ word correction and internet engines. Instead, they increased geocoding matching rate by changing default matching options. Although automated geocoding method by selecting low matching standards has unstable accuracy (Gilboa et al. 2006), the manual geocoding method is still a good way to improve geocoding matching rate.
3.2.2 Data for Assessment

Pesticides Use Report (PUR) and Land Use Maps are usually used to assess pesticides exposure. For instance, in the Gatto et al. (2009) study, the researchers collected commercial application of restricted-use pesticide information since 1990 from the State of California Department of Pesticide Regulation. The restricted-use pesticides are defined in their study as “agents with harmful environmental or toxicological effects”. The PUR location records’ information is according to the Public Land Survey System (PLSS). Each grid of PLSS parcels land is an area of approximately 1 square mile (Gatto et al. 2009). The PUR records should include information about the name of the pesticide’s Active Ingredient (the actual chemical in the product mixture that controls the pest), applied poundage, the type of crop, area of crop field, the method of application, and application date (Gatto et al. 2009).

Land use maps with precise land use information are necessary for pesticide exposure measurement. They can combine with PLSS grid section and describe more details. In Gatto et al. (2009) study, the Land Use Map was obtained from the California Department of Water Resources (CDWR). This data was recorded with “countrywide large-scale surveys of land use and crop cover”.

Sometimes, the Land Use Maps obtained from government agency are not digital maps, which need to be transferred into digital files. Then, the historical digital maps with information about land use and crop type can be combined with PLSS grid section and
PUR to allocate pesticide applications (Gatto et al. 2009).

3.2.3 Ambient Pesticide Exposure Estimates

The buffer function can be used to estimate pesticide exposure in GIS. A buffer is a region around a map feature measured in distance or time (Wade and Sommer. 2006). There are two types of buffers: constant width buffers and variable width buffers. Both of these two types can create a set of feature layers based on attribute values.

In epidemiology studies, the pesticide exposure area is usually defined as an area around a subjects’ home or an area around a pesticide-applied crop field with an appropriate radius. A radius of 500 meters around a subject’s home usually is selected as the exposure scale in spatial epidemiology studies (Chester and Ward 1984; Maccollom et al.1986; McElroy et al. 2003).

Bell et al. (2001) conducted a study which related analyzing agricultural pesticide applications and fetal death. The researchers created a GIS-based method to estimate pesticide exposure. The flow chart (Figure 2-10) shows the procedure of the GIS-based method. Finally, then they calculated the pesticide exposure by using PUR pounds divided by PLSS cells. However, the exposure areas, which were estimated in Bell et al (2001), had a large scale. Their method cannot represent the real exposure of the subjects. For instance, some subjects, who live in the exposure PLSS grid area but far from agriculture area, will count as an exposure. Besides, they assume the pesticides were evenly applied in the PLSS grid cell (Goldberg et al. 2007). But in fact, pesticides were only applied in agriculture
Two years later, Rull et al. (2003) developed a more reasonable method (Figure 3-7) based on Bell et al. (2001) method. In this study, the authors used the buffering function and combined Land use map with the PLSS grid map. They calculated pesticide exposure of each cell using the intersection with the buffered geographic points. Then they summed the exposure of each intersection cell to get the final exposure value. These two additional steps assessed exposure area more accurately and made the results more persuasive.
In 2007, Nuckols et al (2007) improved Rull et al. (2003) measurement method to estimate pesticide exposure (Figure 3-8). They created exposure polygons by using the clip function of GIS. These exposure polygons represented the intersections of crop land and the buffer around each geographic point. The exposure per cell was equal to the application
density per cell (PUR density) divided by land use polygon. Then the total exposure was calculated by adding all exposure values per cell.

Figure 3-9: Flow Chart of Nuckols et al. (2007) Pesticide Exposure Estimate Method (Goldberg et al. 2007)

Nuckols et al. (2007) pesticide exposure estimate method also provides more
comprehensive and accurate ways to measure the exposure of pesticides. Figure 3-9 shows the detail of each procedure. It also can be divided into several steps:

Step 1: Collecting all subjects’ addresses

Step 2: Geocoding all addresses into GIS format map

Step 3: Making 500-m radius buffer of each geographic point to define the zone of potential exposure of pesticide application.

Step 4: Combining PLSS map and land use/land cover map together.

Step 5: Merging the area which contains the buffer and cropping land by using spatial clip function (Figure 3-10).

The CDWR-based pesticide exposure metric (Formula 1) created by Nuckols et al. (2007) provides a more accurate calculation of pesticide exposure. The result shows that CDWR-based pesticide metric can be a better method for pesticide exposure.
\[
EM_k = \sum_{j=1}^{n} \sum_{i=1}^{m} \left( \frac{A_{ij}}{T_{ij}} \right) \times X_{ij},
\]

EM exposure metric for a user-specified pesticide and residence, in pounds

k the pesticide type (active ingredient)

i crop type on which pesticide k was used in Section j (Section is PLSS grid) intersected by the 500-m buffer around the residence

n number of Sections intersected by the 500-m buffer around the residence

m the total number of crop types on which pesticide k was applied in Section j

\(A_{ij}\) the acreage of crop types i within Section j and within 500 m buffer

\(T_{ij}\) total acreage of crop types i within Section j

\(X_{ij}\) total annual pounds of pesticide k applied to crop type i within Section j
3.3 Statistical Software

3.3.1 R Software

The system of R is a powerful statistical computing software for data analysis and graphics. It is created by Ross Ihaka and Robert Gentleman. R language can be seen as the developed S language which was created by Bell Laboratories. R is software as well as a kind of programming language. R, as statistical software, is free to the user under the terms of the Free Software Foundation’s GNU General Public License. There are a lot of functions in R for analyzing data and creating related graphics. The graphic functions can display the graphs in an independent window and save it as many kinds of file types, such as jpg, png, bmp, ps, pdf, emf, pictex, xfig. The results of statistical analysis can be shown directly. Some results in the middle of processing, such as P-value, and regression coefficient, also can be saved in specific files or used for deeply analysis.

The users can continue to analyze many data sets by employing a loop statement. Also, they can combine several different statistical functions into one functional sentence to analyze more complex situations. Most programs using S language can be also used directly in R. There are many sources of that kind of program on the internet. However, R is hard for a non-professional learner to get started.

3.3.2 SAS and SPSS Software

SAS is a popular statistical software used in epidemiological studies. Since a huge pesticides application database needs to be built in the studies which related to accessing
pesticides exposure, in addition, SAS has more functional power and easy to connect with
database. Therefore SAS is very important for this study.

SAS, which was developed in the early 1970s at North Carolina State University, is
the market-leading business intelligence and predictive analytics software. At the
beginning, SAS was used to manage and analyze agricultural field experiments. Now, it
provides powerful statistical and graphical tools for the needs of biology, the social
sciences, engineering, and business. It covers basic analysis and the most complex analyses.
SAS has over 30 years of experience and 43,000 customer sites worldwide.

SAS is mostly used for processing complex collections of data, but it also can be
used for simple math. The techniques used for simple math are also used to develop many
complex changes to any size of data sets. In addition, SAS is more useful when used for
major data manipulation and database processing.

SPSS was developed by three graduate students of Stanford University in 1968. It
can be used for statistical analysis operation, data collection, predictive parsing and
deployment services. SPSS is also a widely used program for statistical analysis in the
social sciences and political sciences.

SPSS can both be used with the Windows point-and-click approach and through
syntax (each has its own benefits, and the user can switch between the approaches). The
data is easily translated into SPSS, and it will reduce the preliminary work which is needed
to explore new data. However, the users have less control of statistical output. And SPSS
also has limited types of data for manipulation.

3.4 Statistical Methods

In epidemiology studies, logistic regression is popularly employed. Logistic regression is a member of the generalized linear models, which include ordinary regression, ANOVA and multivariate statistics such as ANCOVA and log linear regression (Nelder, and Wedderburn. 1972). The type of variables, in epidemiological studies, might be continuous, discrete, dichotomous, or a mix of any of these. Logistic regression allows investigators to predict a discrete outcome from different types of variables. Logistic regression has more useful power, when the independent variables are categorical, or a mix of continuous and categorical.

Prediction of group membership is one function of logistic regression. The outcomes of the analysis are presented with the form of an odds ratio (OR). For instance, in the study of well water consumption and Parkinson’s disease in rural California, Gatto al et. (2009) have employed multivariable unconditional logistic regression methods. They calculated the OR and 95% confidence intervals to predict the relationship between pesticide exposure from well water consumption and PD risk according to multiple variables (age, sex, education, race, family PD history and smoking).
Chapter 4

Methodology

4.1 Study Area

Ohio (Longitude 80°32′W to 84°49′W, Latitude 38°27′N to 41°58′N) covers 44.828 square miles, making it the 34th largest state among US. The following map (Figure 4-1) shows the location of Ohio in US. In addition, it indicates the main cities and surrounding states.

Figure 4-1: Study Area: Ohio Base Map
Source: http://www.netstate.com/states/geography/oh_geography.htm
Figure 4-2 shows the land cover classes in Ohio. Khaki color represents agriculture land and open urban areas (non-impervious, e.g. lawns). The map indicates that most agricultural land is distributed in northwest Ohio. According to the Ohio government’s public website, agriculture is the cornerstone industry of the state, which creates more than 98 billion dollars in revenue. There are more than 77,000 farms in Ohio, and about one seventh of the population is involved in agriculture-related careers.

Figure 4-2: Land Cover of Ohio State
Source: http://home.roadrunner.com/~landcover/Ohio_Landuse_L/Land_Cover_Map.html
Due to the significant position of their agricultural industry, there is extensive farm land in Ohio, which means that pesticide will be applied to farmland in large amounts every year. As a result, one seventh of the state residents will come in contact with agricultural pesticide directly or indirectly. It makes the research on the relationship between pesticide application and PD becomes more essential in this situation.

4.2 Dataset

The PD patients’ dataset in this study is obtained from the Ohio Hospitals Association through Dr Sadik A. Khuder, Professor of Medicine at the University of Toledo. The data was collected from 1999 to 2002, which include 237 PD patients’ ZIP code information in eight counties. These counties are Wood, Lucas, Portage, Montgomery, Mahoning, Hamilton, Franklin and Cuyahoga County.

This dataset contains PD patients’ age, 5-digit ZIP code, city name, county code, county name, primary diagnoses and year of admission. The following Figure 3-3 shows the distribution of the PD patients’ age in the dataset.
Figure 4-3 and table 4-1 above show general information for the age in the dataset, where the maximum age of these patients is 100 and the minimum is 13. The patient who is 13-years old can be considered as genetically predisposed case at early ages. The median value of the data is 62, which means 50 percent of the PD patients are elderly people among the 1699 subjects.

### 4.3 Data Sources

The United States Census Bureau provides the decimal demographic information for the country at various geographical scales, such as block group, ZIP code, county and etc. For example, in this study, the number of total population and the number of total
population of elderly people in ZIP code level can be obtained from the United States Census Bureau website.

Topologically Integrated Geographic Encoding and Referencing (TIGER / TIGER line) is a format to describe geographic attributes such as streets’ shapefiles for geocoding individual addresses. TIGER system has shapefiles of political district boundaries such as block group, city and etc, which can be used for geocoding aggregate data. For instance, a ZIP code layer of Ohio is downloaded to geocode the dataset for this research.

_Pesticides Application Records_

The Ohio Department of Agriculture is in charge of the Ohio agricultural product management. The Pesticide and Fertilizer Regulation Section is one of the Ohio Agriculture Department’s regulatory programs, which provide information about pesticides and fertilizer application regulation and pesticide application licensing for farmers. This program has a pesticide safety education program for private pesticides applicators and requires them to submit pesticide application records for at least three years. It is a good source of pesticide application reports. Figure 3-4 is an example of the form which is required by the Ohio Agriculture Department for private pesticide application records. The information of the pesticides application can be used to calculate total pesticide amount applied for a specific land.
Figure 4-4: Example of Private Pesticide Applicator Restricted Use Recordkeeping Form

4.4 Data Analysis

The population data for year 2000 was obtained from The United States Census Bureau website. The ZIP code, eight-county boundaries’ and Ohio boundary shapefiles were downloaded from TIGER system. Simple descriptive analysis is conducted in eight counties: Wood, Lucas, Portage, Montgomery, Mahoning, Hamilton, Franklin and Cuyahoga County. The first step is extracting the ZIP code areas in the eight counties using clip function. The second step is to geocode the patients’ location. Since the patients’ locations are aggregated at the ZIP code level, in this step, the number of patients in each ZIP code area needs to be calculated and saved as .dbf format. Then, join the table into the attribute table of ZIP code shapefile according to 5-digit ZIP code. Third, join the population data of the 2000 census into ZIP code shapefile’s attribute table. The next step is calculating the incidence rate for each ZIP code area in eight counties by using total patients’ number of each ZIP code area divided by the total population in same area. Finally,
using graduate color map to show the distribution prevalence rate of each ZIP code area.

4.5 Research Methodology

The methods used in this study can be separated into several steps. First, papers about studying the relationship between PD and pesticide exposure and spatial analysis using GIS were downloaded from OhioLink. Second, the methods were extracted into three parts: epidemiology studies, geocoding methods and pesticides exposure assessment method. The techniques which worked well, such as improving matching rate of geocoding or providing more accuracy levels for pesticide exposure estimate were selected for this research. For instance, four types of observational epidemiology study design: cross-sectional, ecologic, cohort and case-control studies, were selected. The geocoding methods in McElroy et al. (2003) and Gilboa et al. study (2006) were selected for representing two examples of geocoding methods, since they can both increase the matching rate. Because the pesticide exposure method in Bell et al. (2001), Rull et al. (2003) and Nuckols et al. (2007) used similar ways to conduct spatial analysis using GIS, they were also selected to represent three different methodologies to estimate pesticide exposure value.

Third, the advantages and disadvantages of these methods were compared in this study. Finally, the dataset was used as an example to express how to conduct spatial analysis of links between Parkinson disease and pesticide exposure. According to the dataset, the ecologic study, aggregated data geocoding method and pesticide exposure
estimates method, which is created based on the study by Nuckols et al. (2007), were selected to express the plan.
Chapter 5

Results

5.1 Comparison of Cross-sectional, Ecologic, Cohort and Case-control Studies

Field epidemiology, cross-sectional, ecologic, case-control and cohort studies are all observational studies. However, most studies which explore the links between pesticide exposure and Parkinson’s disease were case-control or cohort studies. Although all of these can be the options for exploring the association between the hypothesized exposure and disease, however, the approaches of these four methodologies are different.

Table 5-1 shows the advantages and disadvantages of these four study types. The cross-sectional and ecologic studies are less expensive and timesaving than the others, but ecologic study may be limited by “ecologic fallacy”. Cross-sectional and cohort studies are not appropriate for the rare diseases. On the other hand, cohort study is time-consuming and relatively expensive when compared to other studies. However, cross-sectional study allows the researchers to estimate prevalence of all involved factors and study the entire
population. The cohort study can be used to analyze several outcomes related to the exposure. In addition, the incidence rate and relative risks can be calculated in the cohort study and the method can be easily understood. Compared to the cohort study, the case-control study is less expensive and can be completed in a relatively short time. Case-control study can be used to study rare diseases and study multiple risk factors. However, exposure recall bias is a major limitation of their study.
<table>
<thead>
<tr>
<th></th>
<th>cross-sectional studies</th>
<th>Ecologic studies</th>
<th>Cohort studies</th>
<th>Case-control studies</th>
</tr>
</thead>
</table>
| **Advantages**         | 1. Can study entire populations  
2. Can estimate prevalence of all involved factors | 1. Can be implemented in a short time period  
2. less expensive | 1. Can be used to study several outcomes related to the exposure  
2. Provides data to calculate the incidence rate and relative risk  
3. Are clear to understand  
4. Information of all subjects in the study is integrated | 1. Can be used in the study referring to rare diseases  
2. Can be completed over a relatively short time  
3. Less expensive compared with the cohort study  
4. Can be used in the study which refers to multiple risk factors |
| **Disadvantages**      | 1. Not appropriate for rare disease or the disease with short duration  
2. impressionable to misclassification | 1. Can not clarify different characters between individuals  
2. Cannot detect inconspicuous relationships | 1. Cannot be used to study rare diseases.  
2. Expensive  
3. Takes a long time to complete the study | 1. Since the information of the patients is obtained through interview, it might cause incomplete information.  
2. Just can only study one disease.  
3. Complex methods and statistic analysis |

(Adapted from Samet, J M.; Muñoz A. (1998); Armenian, H. K. (1994); and Greenland, S; Robins, J. (1994))
5.2 Comparison of Geocoding Methods

Table 5-2 shows the advantages and disadvantages between geocoding individual address data and geocoding aggregate location data. Geocoding individual address data provides more effective data to conduct spatial analysis. It also gives researchers more flexibility in studying the subjects. But it may take a long time to collect individual address information and the subjects may not want to provide their addresses. In addition, the procedure of geocoding is complex. When some mistakes happen, all procedures need to be reworked. Comparing geocoding individual address data, geocoding aggregate location data is more easily obtained and can be finished in a relatively short time. However, the accuracy level is lower and researchers cannot use the data to deal with distance related issues.

Table 5-2: Advantage and Disadvantage of Geocode Methods

<table>
<thead>
<tr>
<th></th>
<th>Geocoding Individual Address Data</th>
<th>Geocoding Aggregate Location Data</th>
</tr>
</thead>
</table>
| **Advantages**       | 1. Provide high accuracy level of subjects’ location  
                       2. Can calculate distance characteristics | 1. Protect individual privacy  
                       2. Easily obtained  
                       3. Timesaving |
| **Disadvantages**    | 1. Hard to obtain  
                       2. Time-consuming  
                       3. Complex geocode procedure | 1. Low accuracy  
                       2. Cannot compare distance difference between subjects |
| **Appropriate Studies** | Cohort study and case-control study | Ecologic study |
Table 5-3 shows the advantages and disadvantages of geocoding individual data method in McElroy et al.’s study (2003) and Gilboa et al.’s study (2006). The geocoding methods used in both of these studies can improve the final result of the matching rate. The method in McElroy et al. (2003) focuses more on changing human factors, such as using website searching engines and re-contacting subjects. Another advantage is the flexibility, which means the researcher has in changing the geocoding procedure during the study.

Comparing McElroy et al. (2003)’s method to the method used in Gilboa et al. (2006), the latter’s researchers pay more attention to the change of selection of GIS function. This method depends more on GIS software, which means, if any error occurred during the geocoding procedure, the entire process would need to be reworked. Therefore, McElroy et al. (2003) method is more reasonable than Gilboa et al.’s. In addition, McElroy et al. (2003) method allows the researchers get information from subjects directly, which makes the result more accurate. However, Gilboa’s method (2006) also can be a secondary method for geocoding, if most of the subjects were difficult to contact. Since re-contacted subjects need more time and Gilboa’s method (2006) just need change the selection of GIS function, Gilboa’s method (2006) will save much more time than McElroy et al. (2003)’s method.

Table 5-3: Advantage and Disadvantage of Geocoding Individual Data Methods of McElroy et al. (2003) and Gilboa et al. (2006)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Advantages</td>
<td>1. Provides high accuracy</td>
<td>1. Easily conducted</td>
</tr>
<tr>
<td></td>
<td>2. More flexible</td>
<td>2. Time saving</td>
</tr>
<tr>
<td>Disadvantages</td>
<td>1. Time-consuming</td>
<td>1. Low accuracy</td>
</tr>
</tbody>
</table>
5.3 Comparison Methods of Pesticide Exposure Estimates

Table 5-4 compares the advantages and disadvantages among the three methods of pesticide exposure estimates. The method of Bell et al. (2001) is easiest to conduct among these three methods. However, it cannot provide a good accuracy for the study. Comparing Bell et al. (2001)’s method to Rull et al. (2003)’s method is more sensible. It extends and improves Bell et al. (2001)’s method. The researchers divided exposure levels into three tiers, which further improve the accuracy for the results. The method in Nuckols et al (2007)’s study provides the highest accuracy among these three studies. However, the procedure is also most complex and need the most time to be done.

Table 5-4: Comparison Three Pesticide Exposure Estimate Methods

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Advantages</td>
<td>1. Easily conducted</td>
<td>1. relatively higher reasonable than Bell et al. (2001)</td>
<td>1. Can estimate pesticide exposure at a high accuracy level</td>
</tr>
<tr>
<td></td>
<td>2. Divide pesticide exposure into two levels</td>
<td>2. Divide pesticide exposure into three tiers</td>
<td></td>
</tr>
<tr>
<td>Disadvantages</td>
<td>1. Lower accuracy</td>
<td>1. Relatively complex process</td>
<td>1. Complex process</td>
</tr>
</tbody>
</table>
5.4 Comparison of Statistics Software

Table 5-5 indicates the advantages and disadvantages of statistics software of R, SAS and SPSS. R, SAS and SPSS are three popular statistics softwares applied in the different fields. For instance, R is most applied in the study of statistics or simulation in the published papers due to its flexibly in programming. It is also free to download for the public. However, R isn’t appropriate for the non-professional person. SAS is a commercial software, which also can be hard for the beginner. But it can be easily programmed by SQL and connected to database, so SAS is applied popularly in many enterprises due to its powerful functions. SPSS is most applied in the field of economics or geography, which is convenient to input and analyze the data. Nonetheless, researchers have less control of the output and not all types of data can be manipulated in SPSS.

Table 5-5: Advantages and Disadvantages of R, SAS and SPSS Software

<table>
<thead>
<tr>
<th>Statistics Software</th>
<th>R</th>
<th>SAS</th>
<th>SPSS</th>
</tr>
</thead>
</table>
| Advantage           | 1. Flexibly programming  
                      2. Free to download  
                      3. Flexible graph | 1. Powerful functions  
                      2. Formal tables and graphs | 1. Convenient to implement  
                      2. Easy for new starters |
| Disadvantage        | 1. Hard to get started for the non-professional learner | 1. Hard to get started for the non-professional learner  
                      2. Expensive | 1. Less control of statistical output  
                      2. Limited types of data for manipulation |
5.5 Data Analysis

Figure 5-1 shows the locations of the eight counties, which are randomly distributed in Ohio. Figure 5-2 to Figure 5-6 illustrates the distribution of the PD patients’ density in each county. According to theses maps, Mahoning County has the highest prevalence number. Most patients were living in the northeast part of Mahoning County and the ZIP code areas with high incidence rate were distributed around the county’s boundary. Portage County is just next to Mahoning County. These two counties have a similar prevalence situation. Portage County also has a serious incidence rate and most patients were distributed in the county’s middle and southeast sections. Compared with those two counties, Montgomery, Hamilton, Cuyahoga, Lucas and Wood County all have lower number. Franklin County has lowest average of prevalence rate among the eight counties.

![Figure 5-1: Locations of the eight counties of the Dataset](image)
Source: Topologically Integrated Geographic Encoding and Referencing system
Figure 5-2: Distribution of PD Patients’ Density Map in Montgomery County by ZIP code
Source: American FactFinder and Topologically Integrated Geographic Encoding and Referencing system

Figure 5-3: Distribution of PD Patients’ Density Map in Franklin County by ZIP code
Source: American FactFinder and Topologically Integrated Geographic Encoding and Referencing system
Figure 5-4: Distribution of PD Patients’ Density Map in Hamilton County by ZIP code
Source: American FactFinder and Topologically Integrated Geographic Encoding and Referencing system

Figure 5-5: Distribution of PD Patients’ Density Map in Cuyahoga County, Portage County, and Mahoning County by ZIP code
Source: American FactFinder and Topologically Integrated Geographic Encoding and Referencing system
5.5.1 Study Design and Study Unit Selection

Generally, there are two types of spatial data in epidemiologic studies: points (location of cases) and polygonal unit (cluster of cases in a geographic unit) (Matisziw et al, 2008). In previous spatial analytic studies of PD which were discussed before, patients’ addresses were usually selected as the unit of study to ensure the accuracy of the research. Conducting a cohort study is much more expensive and need a long time implement versus the other techniques. In a case-control study, a control group is needed to compare to the case group, however, for this study, there was no patients’ addresses. This made a control selection difficult. As the patients’ locations have been aggregated at ZIP code, city and county level, the ecologic study design is more appropriate for this dataset.

Often protection of personal privacy inhibits the ability to conduct a study. At this time, spatial data aggregation can be used instead to conduct the research. The ZIP code
scale is the smallest area compared to the city and county areas in the dataset. Therefore, using ZIP code scale as the study unit is a better choice for this research. Although the accuracy of the results will be negatively impacted, as the patients’ addresses were missing, it can still provide some insight into the patterns of PD patients’ distribution.

5.5.2 Variables Selection and Calculation

There are four variables, which are critical for this study: the prevalence rate, pesticides exposure, the percentage of elderly people and the agriculture land rate.

*The prevalence rate*

The prevalence rate is the proportion of existing in a population during a specific period. In this study, the prevalence rate is equal to the total number of the PD patients in the study area divided by the total number of population in the same area. The datasets contain eight counties’ PD patients’ information, which includes 273 ZIP code areas. As some ZIP code areas in these counties have no PD patients, these areas need to be found out and set the prevalence rate to zero.

*Pesticides exposure value*

The pesticides exposure value is an index shows the level of people exposed to pesticides. To calculate the pesticide exposure value, the first step is to define the potential exposure area. The assessment method (Figure 4-7) conducted by Nuckols et al. (2007) provided an effective way in calculating pesticide exposure values. However, PD patients’ addresses are a necessary factor to apply Nuckols et al. (2007) method, yet the PD patients’ locations are unavailable in this dataset. Thus an amended method (Figure 4-7) was used to calculate the exposure value in this study. The following chart shows each procedure of this
amended method. First geocoding all patients’ ZIP code to GIS shapefiles. Second, combining land use map and using spatial clip function extract the agriculture land for each ZIP code area. Finally, using formula (2) by combining pesticide application database and agricultural polygons calculate pesticide exposure value.

\[
P_{Ak} = \sum_{i=1}^{m} \left[ \frac{A_i}{T_i} \right] * X_i
\]  

(2)

Figure 5-7: Flow chart of pesticide exposure assessment method

In Nuckols’ study, they calculate the exposure value by using land use section and a 500-meter buffer around the subject’s location. However, ZIP code area is much larger than Nuckols’ study unit. According to the formula used in Nuckols’ study and the data set to be analyzed, the following formula (2) is created to calculate the pesticide exposure for a user-specified pesticide (pesticide application) and the total pesticide applied for the specific ZIP code area is equal to the summation of each type of applied pesticide.
PA pesticide application on a specific pesticide and ZIP code area, in pounds

k the pesticide type (active ingredient)

i crop type on which pesticide k was used in that specific ZIP code area

m the total number of crop types on which pesticide k was applied in that specific ZIP code area

$A_i$ the acreage of crop types i

$T_i$ total acreage of crop types i within that ZIP code area

$X_i$ total annual pounds of pesticide k applied to crop type i

The percentage of elderly people

People over 65 years old are defined as elderly people in this study. The percentage of elderly people can be calculated by using the total population of the zip code area divides total population of elderly people. The United States Census Bureau can provide population database by different types and different area scale.

The agriculture land rate

The agriculture land rate is the ratio of farm land area to the total area of selected zip code areas, which is equal to the total acreage of agriculture land divided by the total acreage of that zip code area.

The land use map can be found from the Ohio Department of Natural Resources’ website. This department provides GIS maps of land use and land cover which contain information of Ohio land surface features. The land use map is updated every year, which
provides a relatively high accuracy of agricultural location for this study. Some universities in Ohio and land management agencies also provide land use maps of Ohio, and they can also be potential sources of land use maps.

5.5.3 Statistical Analysis Procedure

General linear regression model can be used in this study in order to predict the relationship among variables. The data can be analyzed by using a dependent variable $y$: prevalence rate and a number of independent variables $x_1$: pesticide exposure; $x_2$: the percentage of elderly people; $x_3$: the agriculture land rate, $x_{4,5...n}$: different kinds of pesticide exposure that may be related to $y$. and $b_i$ is the coefficient of the variable $x_i$. Then, the general linear regression analysis can be applied to quantify the strength of the relationship between the dependent variables and the independent variables

General linear regression model: $y = b_0 + b_1 x_i, i=1, 2, 3...n$

General linear regression can be applied through the statistic software, such as R and SAS. The hypothesis test will be:

$H_0: b_0 = b_1 = b_2 = b_3 = \cdots = 0$

$H_a: b_0 \neq b_1 \neq b_2 \neq b_3 \neq \cdots \neq 0$

With the significant level $\alpha=0.05$.

By applying the statistic software, the author can discuss the result by analyzing the P-values of each independent variable. Put $x_1$ through $x_i$ into the general linear regression model to determine if there is relationship between the dependent variable $y$ and the independent variable $x_i$. 
Chapter 6

Conclusion and Discussion

6.1 Conclusion

There are a handful of studies about the relationship between PD and pesticides exposure. However, the application of spatial analysis by using GIS can provide more insight to this relationship. GIS software can provide powerful geospatial functions, such as geocoding, buffer, spatial clip and spatial join, for implementing spatial analysis in spatial epidemiology studies. Therefore, this study intends to develop a plan to implement GIS with spatial analysis to explore the association between potential environmental links with Parkinson disease.

In this study, the author not only focus on reviewing, analyzing and evaluating the GIS-based spatial analysis method in previous studies, but also with the idea of “500-meter buffer area = exposure area = ZIP code area”, create a new accessing pesticide exposure method, based on previous ones, to conduct spatial analysis by using ZIP code area as the study unit for the eight counties of Ohio. The results show that the geocoding and pesticide exposure assessment methods are both useful for conducting this study. The geocode methods are selected according to the dataset. In addition, different study method requires different dataset. For example, ecologic study is more appropriate for aggregate dataset. If
the dataset contains individual information, cohort and case-control studies are the well-advised choice.

According to the dataset of this study, since patients’ addresses were missing and the patients’ information was aggregated, the methods need to be amended to adapt the study, by using ecologic study design and ZIP code as the study unit. However, if individual addresses’ information can be obtained, the best method to conduct the study is employing a case-control study. Then use McElroy et al. (2003) geocoding method to geocode the patients’ address. For pesticide exposure assessment section, Nuckols et al. (2007) method can be used to define the exposure area and calculate pesticide exposure value. Finally, all the data can be put into the database and use SAS to manage the data and use general linear regression to analysis the association between pesticide exposure and PD prevalence.

6.2 Future Study

There are some future studies for this research. First, generally, buffer area created by 500-meter radius is much smaller than a regular ZIP code area. But, if some subjects living very close (the distance of each other is smaller than 500m), then the buffer area will overlay each other, the total buffer area may be almost same as a small ZIP code area. In that situation the results might be the same as Nuckols’ pesticide exposure assessing model. Otherwise, it may change the result due to large study unit. Hence, The method of using ZIP code area instead of 500m buffer area needs to be examined in the future work.

Second, in this study, the author just provides the possible sources for pesticide application information. A pesticide application and exposure databases need to be designed and built in the future. A well designed pesticide usage and exposure databases are necessary for conducting the study.
Finally, the climate factors were not considered in this study. But the exposure area may be changed due to some climate factors. For instance, wind blows during pesticide application, the exposure area may be slightly moved according the direction of the wind. Also, if it is raining when applying pesticide, the scale of exposure area will become smaller.
References:


